

Unit 8

Chi Square Tests

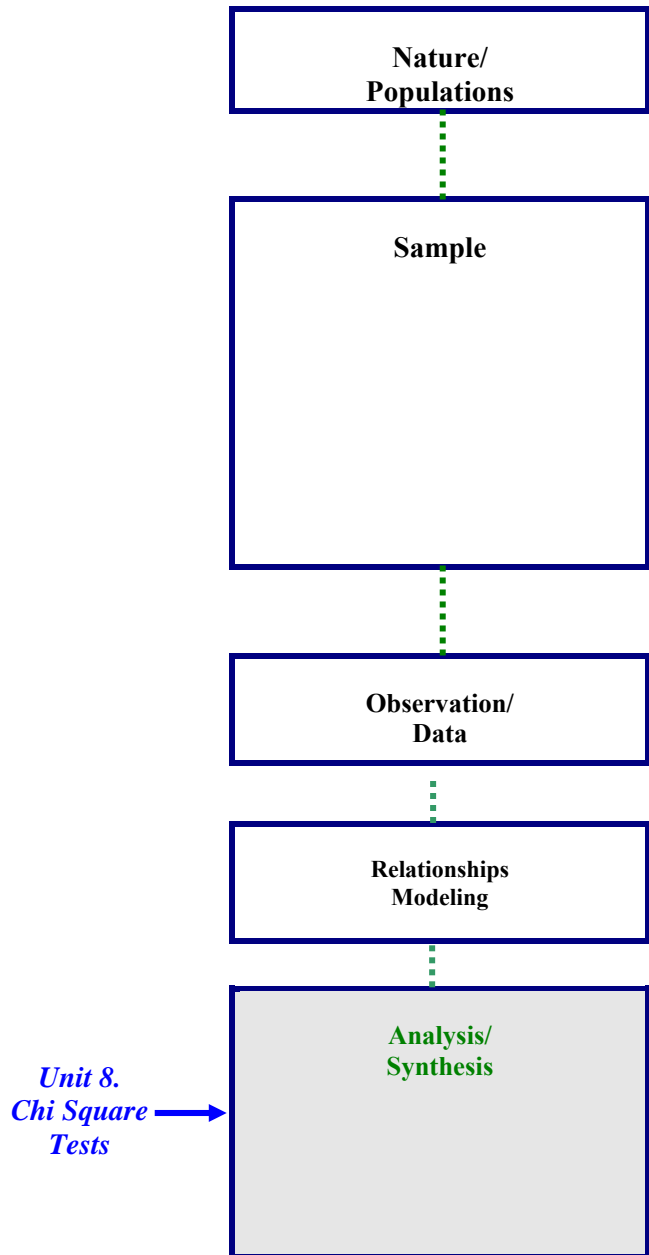
“I shall never believe that God plays dice with the world”

- Albert Einstein (1879-1955)

How many patients died? How many travelers on a cruise ship were exposed to contaminated water? How many will vote for Mitt Romney in 2012? And on and on.... So it goes. This unit is about **counts**.

More to the point, this unit is about **the analysis of counts relative to some “chance model” expectation**. Is the observed count of voters for Mitt Romney in excess of what we might have expected?

1. Unit Roadmap



This unit focuses on the analysis of cross-tabulations of counts called contingency tables. Thus, the data are discrete and whole integer. Examples of count data are number of cases of disease, number of cases of exposure, number of events of voter preference, etc

The structure of a contingency table is a convenient organization of *all the events that could possibly happen*. The contingency table then shows the number of times each “contingency” actually occurred in a given sample. **Example – Suppose there are 2 “contingencies” for disease (yes or no) and 2 “contingencies” for exposure (yes or no). Between disease and exposure, there are 4 possible combinations or “contingencies”.**

The analysis of a contingency table requires a model which predicts the expected counts. Lots of models are possible. The simplest model, and the one described in this unit, is the model of *independence*.

Tip! Chi square tests compare observed counts with null hypothesis model expected counts.

2. Learning Objectives

When you have finished this unit, you should be able to:

- Identify settings where the chi square test is appropriate;
- Explain the equivalence of the null hypotheses of “independence”, “no association”, and equality of proportions;
- Explain the reasoning that underlies the chi square test of “no association”;
- Explain the distinction between “observed” and “expected” counts;
- Calculate, by hand, the chi square test of “no association” for a 2x2 table of observed frequencies ;
- Outline (and perhaps calculate by hand), the steps in a chi square test of no association for an RxC table of observed frequencies;
- Interpret the statistical significance of a chi square test of “no association”.





3. Introduction to Contingency Tables

Contingency table analyses investigate the association between discrete variables that are counts

- Example - Is smoking (yes/no) associated with low birth weight (low/not low)?**
 The number of low birth weight babies born to smokers seems disproportionately high compared to the number of low birth weight babies born to non-smokers? Is this statistically significant?
- Example - Is exposure to lead (yes/no) associated with reduced intelligence (yes/no) in children?**
 The number of lead exposed children with Binet IQ below the cutoff of 85 seems disproportionately great compared to the number of low IQ children who were not exposed to lead.
- Example - Is high income associated with membership in the Republican party?**
 The number of persons with income in the upper 1% who belong to the Republican party seems disproportionately great compared to the number middle income persons who belong to the Republican party.

3a. Contingency Table Counts and Notation

- Example**
 Consider a hypothetical study to investigate the relationship between smoking and impairment of lung function, measured by forced vital capacity (FVC).
- Suppose $n = 100$ people are selected for the study.
- For each person we note their smoking behavior (smoke or don't smoke) and their forced vital capacity, FVC (normal or abnormal). Then we count the number of occurrences of each combination. **Tip!** The contingency table contains counts *not* percentages.

	FVC		
	normal	abnormal	
smoke	 a	 b	a + b
don't smoke	 c	 d	c + d
	a + c	b + d	n = a + b + c + d

these are counts (arrow pointing to cells a and b)

Fixed by sample size (arrow pointing to the total n)

- One scenario is the following set of counts

	fvc		
	abn	normal	
smoke	50	0	50
don't smoke	0	50	50
	50	50	100

What can be said about the relationship between fvc and smoking?

- All 50 smokers have an abnormal FVC
- And all 50 non-smokers have normal FVC
- This is an illustration of a **perfect association**: *Once smoking status is known, FVC status is known also.*

- Another scenario is the following set of counts

	fvc		
	abn	normal	
smoke	25	25	50
don't smoke	25	25	50
	50	50	100

- In this scenario, half (25) of the smokers have an abnormal FVC
- But, also, half (25) of the 50 non-smokers have an abnormal FVC.
- This is an illustration of **no association**: *Knowledge of smoking, one way or the other, does not help in predicting FVC status.*
- Here, “no association” is saying: *Lung function, as measured by FVC, is independent of smoking status.*

Introduction to observed versus expected counts.

- **Observed** counts are represented using the notation “**O**” or “**n**”.
- **Expected** counts are the null hypothesis expected counts. They are represented using the notation “**E**”

		<u>FVC</u>		
		Abnormal	Normal	
Smoke	O_{11}	O_{12}		$O_{1.}$
Don't smoke	O_{21}	O_{22}		$O_{2.}$
	$O_{.1}$	$O_{.2}$		$O_{..}$

How to read the “O” notation and its subscripts -

O_{21} = count in the cell that is in row “2” and column “1”

O_{21}	
The first subscript tells you the “row” Example: O_{21} is a cell count in row “2”	The second subscript tells you the “column” Example: O_{21} is a cell count in column “1”

How to read subscripts that are dots-

A dot subscript references a total, either a row total or a column total or both.

$O_{2.}$	$O_{.1}$	$O_{..}$
$O_{2.}$ is the row “2” total. It is taken over all the columns	$O_{.1}$ is the column “1” total. It is taken over all the rows	$O_{..}$ is the “grand” total. It is taken over all rows and all columns

- **Example:** Here are the **observed** counts in another scenario

		<u>FVC</u>		
		Abnormal	Normal	
Smoke		$O_{11}=40$	$O_{12}=10$	$O_{1.}=50$
Don't smoke		$O_{21}=5$	$O_{22}=45$	$O_{2.}=50$
		$O_{.1}=45$	$O_{.2}=55$	$O_{..}=100$

- $O_{21} = 5$ is # in row 2 column 1
- $O_{12} = 10$ is # in row 1 column 2
- $O_{1.} = 50$ is the row 1 total
- $O_{.1} = 45$ is the column 1 total

In the next section, we'll learn about the expected counts "E"

You will see that "expected" counts are the null hypothesis counts that would have been expected to occur under the assumption that the null hypothesis is true.

3b. Contingency Table Counts and Degrees of Freedom

In a contingency table, the focus is on the *distribution of counts among the various "contingencies"*

The row and column totals are **fixed**.

In this context, the **"degrees of freedom"** are the number of individual cell counts that are **free to vary**:

- **Example - 2x2 table**

\textcircled{x}	$n_1 - x$	n_1	∴ we have "freedom" to fill in only one of the cells
$n_3 - x$	$n_2 - (n_3 - x)$	n_2	
n_3	n_4	n	⇒ 1 degree of freedom

• **Examples larger tables**

x	x	

= 2 d.f.

x	x	x	

= 3 d.f.

x	x	
x	x	

= 4 d.f.

x	x	x	x	

= 4 d.f.

Tip! In each scenario, the last column is not free and the last row is not free.

Degrees of Freedom
R x C table
General Test of No Association

= (#rows – 1) x (#columns – 1)

= (R – 1)(C – 1)

4. Introduction to the Contingency Table Hypothesis Test of No Association

Recall from Unit 7 (*Hypothesis Testing*) the steps we followed to develop a “proof by contradiction” approach to hypothesis tests.

Steps in Hypothesis Testing

1. Identify the research question.
2. State the null hypothesis assumptions necessary for computing probabilities.
3. Specify H_0 and H_A .
4. “Reason” an appropriate test statistic.
5. Specify an “evaluation” rule.
6. Perform the calculations.
7. “Evaluate” findings and report.
8. Interpret in the context of biological relevance.
9. (Accompany the procedure with an appropriate confidence interval)

Example for Illustration:

Suppose the following were *observed* in the investigation of smoking and forced vital capacity.

		<u>FVC</u>		
		Abnormal	Normal	
Smoke	$O_{11}=40$	$O_{12}=10$	$O_{1.}=50$	
Don't smoke	$O_{21}=5$	$O_{22}=45$	$O_{2.}=50$	
	$O_{.1}=45$	$O_{.2}=55$	$O_{..}=100$	

Nature _____ Population/ Sample _____ Observation/ Data _____ Relationships/ Modeling _____ Analysis/ Synthesis

1. Identify the research question.-

Is smoking associated with impaired lung function, measured by forced vital capacity (FVC)?

2. State the null hypothesis assumptions necessary for computing probabilities.-

The “nothing interesting is going on” statement that defines the null hypothesis here is the following: There is *no association* between smoking and impaired lung function as measured by forced vital capacity (FVC).

3. Specify H_0 and H_A .

Let

π_1 = the proportion of smokers with abnormal fvc

π_2 = the proportion of non-smokers with abnormal fvc

Under the null hypothesis assumption, the proportion of persons with abnormal fvc *is the same*, regardless of smoking status.

$$H_0: \pi_1 = \pi_2$$

Whereas, when the alternative hypothesis is true, the proportion of persons with abnormal fvc *will be different*, depending on smoking status.

$$H_A: \pi_1 \neq \pi_2$$

4. Reason an appropriate test statistic.

The appropriate statistic here compares the observed counts “O” to the null hypothesis expected counts “E”.

How to Solve for the Null Hypothesis Expected Counts E

The reasoning proceeds as follows.

(1) When the null hypothesis is true

- $\pi_1 = \pi_2 = \pi$ where π is the common (*null hypothesis*) value

Nature _____ Population/ Sample _____ Observation/ Data _____ Relationships/ Modeling _____ Analysis/ Synthesis

(2) But now we need a guess of the common π

- The common π is estimated as the observed overall proportion of abnormal fvc.

$$\hat{\pi} = \frac{45}{100} = \frac{\text{column 1 total}}{\text{grand total}}, \text{ or a bit more formally ...}$$

$$\hat{\pi} = \frac{O_{11} + O_{21}}{O_{11} + O_{12} + O_{21} + O_{22}} = \frac{O_{.1}}{O_{..}} = \frac{40 + 5}{100} = 0.45$$

(3) Next, assume π_1 and π_2 are equal to the same null hypothesis estimate $\hat{\pi} = 0.45$

Thus, under the assumption that H_0 is true (meaning *no association, independence*), the proportion with abnormal fvc among smokers as well as among non-smokers should be the same as in the overall population, that is,

$$\pi_{1;\text{null}} = \pi_{2;\text{null}} = \hat{\pi} = 0.45$$

(4) Compute the null hypothesis **expected counts of abnormal fvc** in each of the two groups

Under the null hypothesis we expect 45% of the 50 smokers, or 22.5 persons, to have abnormal fvc. We also expect 45% of the 50 non-smokers, or 22.5 persons, to have abnormal fvc.

TIP!! These expected counts are NOT whole integers. That's okay. **Do NOT round.**

$$\text{Expected \# smokers w abnormal FVC} = (\#\text{Smokers})(\hat{\pi}) = (50)(.45) = 22.5 = E_{11}$$

$$\text{Expected \# NONsmokers w abnormal FVC} = (\#\text{NONSmokers})(\hat{\pi}) = (50)(.45) = 22.5 = E_{21}$$

(5) Compute the null hypothesis **expected counts of normal fvc** in each of the two groups

We get this by subtraction since the numbers of smokers and non-smokers are fixed!

Under the null hypothesis we expect 55% of the 50 smokers, or 27.5 persons, to have normal fvc. Similarly, we also expect 55% of the 50 non-smokers, or 27.5 persons, to have normal fvc.

Thus the following **null hypothesis expected counts “E”** emerge.

		<u>FVC</u>		
		Abnormal	Normal	
Smoke	E ₁₁ =22.5	E ₁₂ =27.5	E _{1.} =50	
Don't smoke	E ₂₁ =22.5	E ₂₂ =27.5	E _{2.} =50	
		E _{.1} =45	E _{.2} =55	E _{..} =100

- E₂₁=22.5 E₁₂=27.5
- E_{1.}=50 E_{.1}=45

Note -

- The expected row totals match the observed row totals.
- The expected column totals match the observed column totals.
- These totals have a special name - “marginals”.
- The “marginals” are treated as fixed constants (“givens”).

The appropriate test statistic is a chi square statistic, *provided the sample sizes are sufficiently large*
 The chi square statistic here is a comparison of observed and null hypothesis expected counts.

$$\begin{aligned}
 \text{Chi Square}_{df} = \chi_{df}^2 &= \sum_{\text{all cells "i,j"}} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] \\
 &= \sum_{\text{all cells "i,j"}} \left[\frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}} \right]
 \end{aligned}$$

5. Specify an Evaluation Rule.

A closer look at the chi square statistic suggests the following:

When the null hypothesis is true, the differences ($O - E$) will tend to be small.

The resulting chi square statistic will tend to have a value that is small

But when the alternative hypothesis is true, then at least some of the differences ($O - E$) will be large

The resulting chi square statistic will tend to have a value that is positive, large.

The development of an evaluation rule follows the same approach as what we learned in Unit 7 (Hypothesis Testing). We begin by assuming the null hypothesis is true and then calculate the null hypothesis chances of the chi square statistic being as extreme as, or more extreme than, the value obtained for our data.

2 x 2 Table
Chi Square Test of No Association
for sufficiently large sample size

$$\text{Chi Square}_{df=1} = \chi_{df=1}^2 = \sum_{\text{all cells "i,j"}} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

$$= \sum_{\text{all cells "i,j"}} \left[\frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}} \right]$$

Rejection of the null hypothesis occurs for large values of the chi square statistic and accompanying small p-values

6. Perform the Calculations.

Recall the observed and null hypothesis expected counts.

Observed Counts, “O”

		<u>FVC</u>		
		Abnormal	Normal	
Smoke		$O_{11}=40$	$O_{12}=10$	$O_{1.}=50$
Don't smoke		$O_{21}=5$	$O_{22}=45$	$O_{2.}=50$
		$O_{.1}=45$	$O_{.2}=55$	$O_{..}=100$

Null Hypothesis Expected Counts, “E”

		<u>FVC</u>		
		Abnormal	Normal	
Smoke		$E_{11}=22.5$	$E_{12}=27.5$	$E_{1.}=50$
Don't smoke		$E_{21}=22.5$	$E_{22}=27.5$	$E_{2.}=50$
		$E_{.1}=45$	$E_{.2}=55$	$E_{..}=100$

$$\text{Chisquare}_{DF=1} = \left[\frac{(40-22.5)^2}{22.5} \right] + \left[\frac{(10-27.5)^2}{27.5} \right] + \left[\frac{(5-22.5)^2}{22.5} \right] + \left[\frac{(45-27.5)^2}{27.5} \right]$$

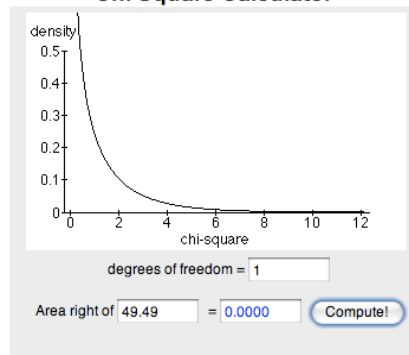
$$= 49.4949$$

P-Value Calculation

P-value = probability [chi square_{DF=1} ≥ 49.4949]

<<<<.0001

Chi Square Calculator



<http://www.stat.tamu.edu/~west/applets/chisqdemo.html>

7. Evaluate Findings and Report.

Under the null hypothesis assumption of no association of smoking with abnormal forced vital capacity, the chances of obtaining a chi square statistic as large as 49.40 or greater were less than 1 chance in 10,000. Thus, the assumption of the null hypothesis, when examined in light of the data, has led to an extremely unlikely conclusion. → *Reject the null hypothesis.*

The data, as given, suggests an association. Further analyses are needed to understand its nature.

How large is large enough?

What are the sample size requirements?

When can I use this test?

This chi square test assumes that the sample size is “sufficiently large”. But what is that?

Different texts and sources suggest different “rules of thumb”. They’re similar.

- * Do NOT use the chi square test if the total sample size is less than 20. Instead, use the Fisher’s Exact Test (discussed in PubHlth 640)
- * Do NOT use the chi square test if $20 < n < 40$ and any individual null hypothesis expected cell count is less than 5. Instead use the Fisher’s Exact test.
- * When the sample size is > 40 , expected cell counts that are as small as 1 or 2 can be tolerated.

5. The Chi Square Test of No Association in an R x C Table

The general test of no association for a 2x2 table is easily extended to a general test of no association for an RxC table

- For one cell, when the null hypothesis is true,

$$\frac{\left[\begin{array}{cc} \text{Observed} & \text{Expected} \\ \text{Count} & - \text{Count} \end{array} \right]^2}{\text{Expected Count}}$$

is distributed Chi Square (df = 1) approximately.

- Summed over all cells in an R x C table, when the null hypothesis is true,
In a table that has “R” rows and “C” columns, the same calculation is repeated RC times and then summed to obtain

R x C Table
Chi Square Test of No Association
for sufficiently large sample size

$$\text{Chi Square Statistic}_{DF=(R-1)(C-1)} = \sum_{i=1}^R \sum_{j=1}^C \frac{\left[\begin{array}{cc} \text{Observed} & \text{Expected} \\ \text{Count (i,j)} & - \text{Count (i,j)} \end{array} \right]^2}{\text{Expected Count (i,j)}}$$

Degrees of Freedom = DF = (R-1) (C-1)

Rejection of the null hypothesis occurs for large values of the chi square statistic and accompanying small p-values

- This chi square test statistic is distributed Chi Square (df = [R-1][C-1]) approximately when the null hypothesis is true.

Example

Suppose we wish to investigate whether or not there is an association between income level and how regularly a person visits his or her doctor. Consider the following count data.

Income	Last Consulted Physician			Total
	< 6 months	7-12 months	>12 months	
< \$6000	186	38	35	259
\$6000-\$9999	227	54	45	326
\$10,000-\$13,999	219	78	78	375
\$14,000-\$19,999	355	112	140	607
> \$20,000	653	285	259	1197
Total	1640	567	557	2764

Notation for Observed (“O” or “n”) Counts in the RxC Setting:

		Columns, “j”			
		j = 1	...	j = C	
Rows, “i”	i = 1	$O_{11}=n_{11}$...	$O_{1C}=n_{1C}$	$N_{1.} = O_{1.}$
			
	i = R	$O_{R1}=n_{R1}$...	$O_{RC}=n_{RC}$	$N_{R.} = O_{R.}$
		$N_{.1} = O_{.1}$...	$N_{.C} = O_{.C}$	$N = O_{..}$

Definition of the π_{ij} in the RxC Setting:

π_{ij} = the probability of having income level “i” and elapsed consult time “j”
 EG - π_{11} = probability [income is <\$6000 AND time since last visit is \leq 6 mos]

$\pi_{i.}$ = the overall (marginal) probability that income is at level “i”
 EG: $\pi_{1.}$ = probability [income is <\$6000]

$\pi_{.j}$ = the overall (marginal) probability that time since last visit is at level “j”
 EG: $\pi_{.1}$ = probability [time since last visit is \leq 6 months]

Review of independence in the tossing of two independent coins

Recall the example of tossing a fair coin two times. Under independence, we learned that

$$\Pr [\text{“heads” on toss 1 } \underline{\text{and}} \text{ “heads” on toss 2 }] = (.50)(.50) = .25$$

Let

$$\begin{aligned} \pi_{1.} &= \text{Probability of “heads” on toss 1, regardless of outcome of toss 2} \\ \pi_{.2} &= \text{Probability of “heads” on toss 2, regardless of outcome on toss 2} \end{aligned}$$

Now let

$$\pi_{12} = \text{Probability of “heads” on toss 1 and “heads” on toss 2}$$

Independence →

$$\begin{aligned} \pi_{12} &= [\text{probability heads on toss 1}] \times [\text{probability heads on toss 2}] \\ &= [\pi_{1.}] [\pi_{.2}] \end{aligned}$$

Thus, under independence

$$\begin{array}{ccc} \pi_{ij} & = & [\pi_{i.}] [\pi_{.j}] \\ \downarrow & & \downarrow \quad \downarrow \\ \Pr [\text{“i” x “j” combination}] & = & [\text{Marginal “i” prob}] \times [\text{Marginal “j”}] \end{array}$$

Application of Independence to the RxC Setting: The income x consult time example

Let

$$\begin{aligned} \pi_{1.} &= \text{Probability that income is } < \$6000, \text{ overall} \\ \pi_{.1} &= \text{Probability that consult time is } \leq 6 \text{ months, overall} \end{aligned}$$

Now let

$$\pi_{11} = \Pr [(\text{income} < \$600) \text{ and } (\text{consult time } \leq 6 \text{ months})]$$

Independence →

$$\begin{aligned} \pi_{11} &= \Pr[\text{income} < \$6000] * \Pr [\text{consult time } \leq 6 \text{ months}] \\ &= \pi_{1.} * \pi_{.1} \text{ That is,} \end{aligned}$$

$$\pi_{11} = (\pi_{1.})(\pi_{.1}) \text{ under independence}$$



Example, continued-

π_i = Probability that income is level “i”

π_j = Probability that time since last visit is at level “j”

π_{ij} = Probability income is level “i” AND time since last visit is at level “j”

Under Independence,

$$\pi_{ij} = [\pi_i] [\pi_j]$$

Null Hypothesis Assumptions for RxC General Chi Square Test of NO Association

1. The contingency table of count data is a random sample from some population
2. The cross-classification of each individual is independent of the cross-classification of all other individuals.

Specify Null and Alternative Hypotheses

$$H_O : \pi_{ij} = \pi_i \pi_j$$

$$H_A : \pi_{ij} \neq \pi_i \pi_j$$

Reason an Appropriate Test Statistic

We need to solve for the null hypothesis expected counts. To do this, we need the null hypothesis probabilities. These are obtained as follows.

$$\hat{\pi}_{ij} = \hat{\pi}_i \hat{\pi}_j \text{ by independence and where}$$

$$\hat{\pi}_i = \frac{n_{i.}}{n} = \frac{\text{row "i" total}}{\text{grand total}}$$

$$\hat{\pi}_j = \frac{n_{.j}}{n} = \frac{\text{column "j" total}}{\text{grand total}}$$

Null Hypothesis Expected Counts E_{ij}

$$E_{ij} = (\# \text{ trials})[\hat{\pi}_{ij} \text{ under null}] = (n)\hat{\pi}_i \hat{\pi}_j = \frac{[\text{row "i" total}][\text{column "j" total}]}{n}$$

Specify an Evaluation Rule/Test Statistic

The reasoning is the same as that for the 2x2 table test of general association. For each cell, the comparison of the observed versus null hypothesis expected counts is obtained using:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The **chi square test statistic of general association** is, again, the sum of these over all the cells in the table:

$$\text{Chi Square Statistic}_{DF=(R-1)(C-1)} = \sum_{i=1}^R \sum_{j=1}^C \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

Behavior of the Test Statistic under the assumption of the null hypothesis

When the null hypothesis is true,

$$\text{Chi Square Statistic} = \sum_{i=1}^R \sum_{j=1}^C \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] \text{ is distributed } \chi_{df=(R-1)(C-1)}^2$$

The null hypothesis is rejected for large values of the test statistic. Thus, evidence for rejection of the null hypothesis is reflected in the following (all will occur)

- LARGE value of test statistic
- SMALL value of achieved significance (p-value)
- Test statistic value that EXCEEDS CRITICAL VALUE threshold

Perform the Calculations

(1) For each cell, compute the expected cell count under the assumption of independence

$$E_{ij} = \frac{[\text{row "i" total}][\text{column "j" total}]}{n}$$

(2) For each cell, compute

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Example, continued -

Observed Counts (*this is just the table on page 18 again with the “O” notation provided*)

Last Consulted Physician

Income	≤ 6 months	7-12 months	>12 months	Total
< \$6000	O ₁₁ = 186	O ₁₂ =38	O ₁₃ =35	O _{1.} =259
\$6000-\$9999	O ₂₁ =227	O ₂₂ =54	O ₂₃ =45	O _{2.} =326
\$10,000-\$13,999	O ₃₁ =219	O ₃₂ =78	O ₃₃ =78	O _{3.} =375
\$14,000-\$19,999	O ₄₁ =355	O ₄₂ =112	O ₄₃ =140	O _{4.} =607
≥ \$20,000	O ₅₁ =653	O ₅₂ =285	O ₅₃ =259	O _{5.} =1197
Total	O _{.1} =1640	O _{.2} =567	O _{.3} =557	O _{..} =2764

Null Hypothesis Expected Counts – *note that each entry is (row total)(column total)/(grand total)*

Last Consulted Physician

Income	≤ 6 months	7-12 months	>12 months	Total
< \$6000	E ₁₁ = $\frac{(259)(1640)}{2764} = 153.68$	E ₁₂ =53.13	E ₁₃ =52.19	E _{1.} =259
\$6000-\$9999	E ₂₁ =193.43	E ₂₂ =66.87	E ₂₃ =65.70	E _{2.} =326
\$10,000-\$13,999	E ₃₁ =222.50	E ₃₂ =76.93	E ₃₃ =75.57	E _{3.} =375
\$14,000-\$19,999	E ₄₁ =360.16	E ₄₂ =124.52	E ₄₃ =122.32	E _{4.} =607
≥ \$20,000	E ₅₁ =710.23	E ₅₂ =245.55	E ₅₃ = $\frac{(1197)(557)}{2764} = 241.22$	E _{5.} =1197
Total	E _{.1} =1640	E _{.2} =567	E _{.3} =557	E _{..} =2764

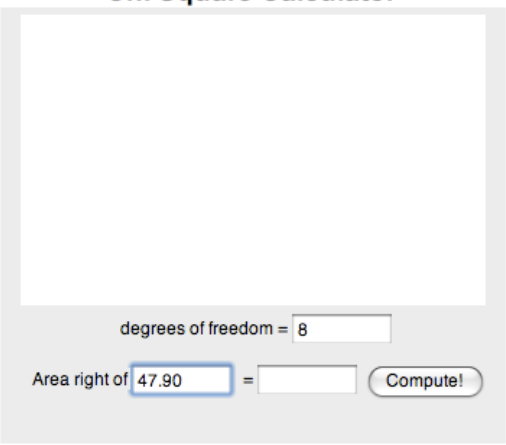
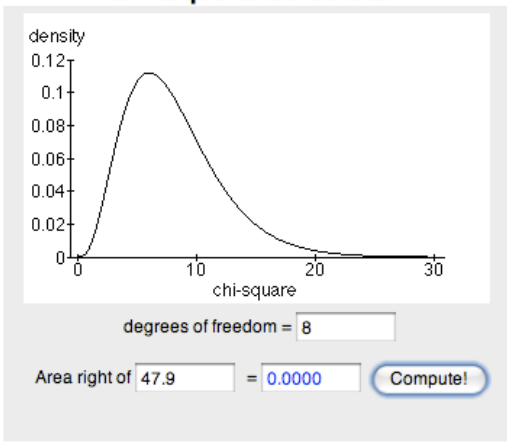
Nature _____ Population/ Sample _____ Observation/ Data _____ Relationships/ Modeling _____ Analysis/ Synthesis

$$\chi^2_{(R-1)(C-1)} = \sum_{\text{all cells}} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] = \frac{(186 - 153.68)^2}{153.68} + \dots + \frac{(259 - 241.22)^2}{241.22} = 47.90$$

with degrees of freedom = (R-1)(C-1) = (5-1)(3-1) = 8

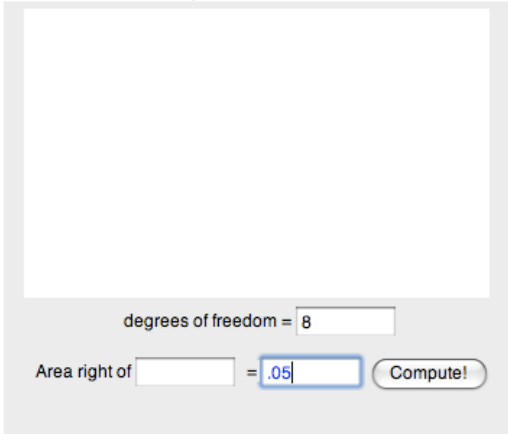
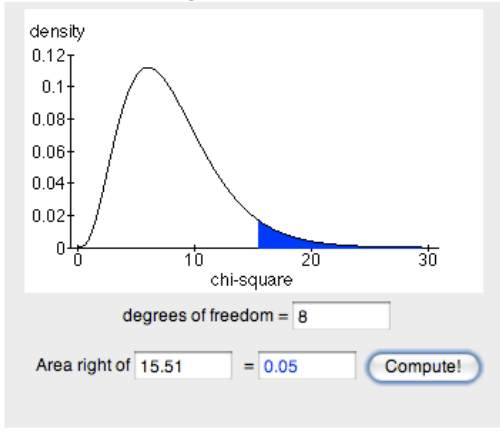
P-value Calculation

p-value = Probability [Chi square with df=8 \geq 47.90] \ll .0001
 Such an extremely small p-value is statistically significant \rightarrow
 Reject the null hypothesis.

	
<p>Enter “8” for degrees of freedom. Enter test statistic value 47.90 Click compute</p>	<p>Calculator will return 0.0000 Thus, p-value \ll 0.0001</p>

<http://www.stat.tamu.edu/~west/applets/chisqdemo.html>

Statistical Decision Using Critical Region (type 1 error = 0.05) Approach

	
<p>Enter “8” for degrees of freedom. Enter area under the curve = .05 Click compute</p>	<p>Calculator will return 15.51 Thus, the .05 critical value of the statistic is 15.51</p>

<http://www.stat.tamu.edu/~west/applets/chisqdemo.html>

$\chi^2_{.95;df=8} = 15.51$ is our critical value.

Observed statistic = 47.90 >> $\chi^2_{.95;df=8} = 15.51 \rightarrow$

Reject the null hypothesis.

Evaluate Findings and Report -

Under the null hypothesis assumption of no association of “time since last visit with a physician” and “income”, the chances of obtaining a chi square statistic with 8 df as large as 47.90 or greater were less than 1 chance in 10,000. Thus, the assumption of the null hypothesis, when examined in light of the data, has led to an extremely unlikely conclusion. \rightarrow **Reject the null hypothesis.**

Thus, these data provide statistically significant evidence that time since last visit to the doctor is NOT independent of income, that there is an association between income and frequency of visit to the doctor.

Important note! What we’ve learned is that there **is** an association, but **not its nature**. This will be considered further in PubHlth 640, *Intermediate Biostatistics*.

Nature _____ Population/ Sample _____ Observation/ Data _____ Relationships/ Modeling _____ Analysis/ Synthesis

6. (For Epidemiologists) Special Case: More on the 2x2 Table

Sometimes, a “a, b, c, d” notation is used for a 2x2 table

Many epidemiology texts use a different notation for representing the counts in a 2x2 table. The counts are “a”, “b”, “c”, and “d” as follows.

		2nd Classification Variable		
		1	2	
1st Classification	1	a	b	a + b
	2	c	d	c + d
		a + c	b + d	n

The “O” and “E” formula for the test of no association in a 2x2 table

$$\chi^2_{DF=1} = \sum_{i=1}^2 \sum_{j=1}^2 \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

The “a,b,c,d, and n” formula for the test of no association in a 2x2 table

$$\chi^2_{DF=1} = \frac{n(ad - bc)^2}{(a+c)(b+d)(c+d)(a+b)}$$

7. Hypotheses of Independence or No Association

“Independence”, “No Association”, “Homogeneity of Proportions” are alternative wordings for the same thing.

Example,

- (1) “Length of time since last visit to physician” is independent of “income” means that income has no bearing on the elapsed time between visits to a physician. The expected elapsed time is the same regardless of income level.
- (2) There is no association between coffee consumption and lung cancer means that an individual’s likelihood of lung cancer is not affected by his or her coffee consumption.
- (3) The equality of probability of success on treatment (experimental versus standard of care) in a randomized trial of two groups is a test of homogeneity of proportions.

The hypotheses of “independence”, “no association”, “homogeneity of proportions” are equivalent wordings of the same null hypothesis in an analysis of contingency table data.

Nature _____ Population/ Sample _____ Observation/ Data _____ Relationships/ Modeling _____ Analysis/ Synthesis

Appendix
Relationship Between the Normal(0,1) and the Chi Square Distributions
For the interested reader

This appendix explains how it is reasonable to use a continuous probability model distribution (the chi square) for the analysis of discrete (counts) data, in particular, investigations of association in a contingency table.

- Previously (see Unit 6, *Estimation*), we obtained a chi square random variable when working with a function of the sample variance S^2 .
- It is also possible to obtain a chi square random variable as the square of a Normal(0,1) variable. **Recall that this is what we have so far ...**

IF	THEN	Has a Chi Square Distribution with DF =
Z has a distribution that is Normal (0,1)	Z^2	1
X has a distribution that is Normal (μ, σ^2), so that Z - score = $\frac{X - \mu}{\sigma}$	$\{ Z\text{-score} \}^2$	1
X_1, X_2, \dots, X_n are each distributed Normal (μ, σ^2) and are independent, so that \bar{X} is Normal ($\mu, \sigma^2/n$) and Z - score = $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$\{ Z\text{-score} \}^2$	1
X_1, X_2, \dots, X_n are each distributed Normal (μ, σ^2) and are independent and we calculate $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$	$\frac{(n - 1)S^2}{\sigma^2}$	(n-1)

Our new formulation of a chi square random variable comes from working with a Bernoulli, the sum of independent Bernoulli random variables, and the central limit theorem. What we get is a great result. The chi square distribution for a continuous random variable can be used as a good model for the analysis of discrete data, namely data in the form of counts.

	<p>Z_1, Z_2, \dots, Z_n are each Bernoulli with probability of event = π.</p> $E[Z_i] = \mu = \pi$ $\text{Var}[Z_i] = \sigma^2 = \pi(1 - \pi)$ <p style="text-align: center;">↓</p>	
	<p>1. The net number of events $X = \sum_{i=1}^n Z_i$ is Binomial (N, π)</p> <p>2. We learned previously that the distribution of the <u>average</u> of the Z_i is well described as Normal($\mu, \sigma^2/n$).</p> <p style="text-align: center;">Apply this notion here: By convention,</p> $\bar{Z} = \frac{\sum_{i=1}^n Z_i}{n} = \frac{X}{n} = \bar{X}$ <p style="text-align: center;">↓</p>	
	<p>3. So perhaps the distribution of the <u>sum</u> is also well described as Normal. At least approximately</p> <p style="padding-left: 40px;">If \bar{X} is described well as Normal ($\mu, \sigma^2/n$)</p> <p style="padding-left: 40px;">Then $X = n\bar{X}$ is described well as Normal ($n\mu, n\sigma^2$)</p> <p style="text-align: center;">↓</p>	
	<p style="text-align: center;">Exactly: X is distributed Binomial(n, π)</p> <p style="text-align: center;">Approximately: X is distributed Normal ($n\mu, n\sigma^2$)</p> <p style="text-align: center;">Where: $\mu = \pi$ and $\sigma^2 = \pi(1 - \pi)$</p>	

Putting it all together ...

IF	THEN	Comment
<p>X has a distribution that is <u>Binomial</u> (n,π) <u>exactly</u></p>	<p>X has a distribution that is <u>Normal</u> (nμ, nσ²) <u>approximately</u>, where</p> $\mu = \pi$ $\sigma^2 = \pi(1-\pi)$ <p style="text-align: center;">↓</p>	
	$Z\text{-score} = \frac{X - E(X)}{SD(X)}$ $= \frac{X - n\mu}{\sqrt{n\sigma}}$ $= \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}}$ <p>is approx. Normal(0,1)</p> <p style="text-align: center;">↓</p>	
	<p>{ Z-score }² has distribution that is well described as Chi Square with df = 1.</p>	<p>We arrive at a continuous distribution model (chi square) approximation for count data.</p>

Thus, the $\{Z\text{-score}\}^2$ that is distributed approximately Chi Square (df=1) is the $(O-E)^2/E$ introduced previously.

- **Preliminaries**

$$X = \text{“Observed”} = O$$

$$n\pi = \text{“Expected”} = E$$

- **As n gets larger and larger**

$$n\pi(1-\pi) \rightarrow n\pi(1) = \text{“Expected”} = E$$

- **Upon substitution,**

$$\{Z\text{-Score}\}^2 = \left\{ \frac{X-n\pi}{\sqrt{n\pi(1-\pi)}} \right\}^2 \rightarrow \left\{ \frac{X-n\pi}{\sqrt{n\pi(1)}} \right\}^2 = \left\{ \frac{O-E}{\sqrt{E}} \right\}^2 = \frac{(O-E)^2}{E}$$

Thus,

- For **one cell**, when the *null hypothesis is true*, the **central limit theorem** gives us

$$\frac{\left[\begin{array}{cc} \text{Observed} & \text{Expected} \\ \text{Count} & - \text{Count} \end{array} \right]^2}{\text{Expected Count}} \text{ is Chi Square (df = 1) approximately.}$$

- For **RC cells**, when the *null hypothesis is true*, the **central limit theorem** and the **definition of the chi square distribution** give us

$$\sum_{i=1}^R \sum_{j=1}^C \frac{\left[\begin{array}{cc} \text{Observed} & \text{Expected} \\ \text{Count}_{ij} & - \text{Count}_{ij} \end{array} \right]^2}{\text{Expected Count}_{ij}} \text{ is Chi Square [df=(R-1)(C-1)] approx.}$$

Nature _____ Population/ Sample _____ Observation/ Data _____ Relationships/ Modeling _____ Analysis/ Synthesis