Topic 8

Chi Square Tests

Topics

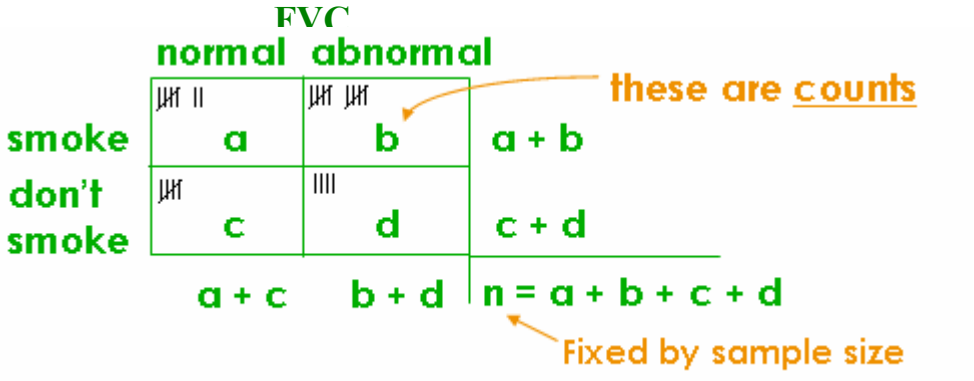
1. Introduction to Contingency Tables

We wish to Explore the Association Between Two Variables, Both Discrete

- **Example - Is smoking associated with low birth weight?**
- **A common goal of many research studies is investigation of the association of 2 factors, both discrete; eg – smoking (yes/no) and low birth weight (yes/no).**
- **In Topic 8, our focus is in the setting of two categorical variables, such as smoking and low birth weight, and the use of chi-square tests of association and homogeneity.**

Introduction to Contingency Tables

- **Example - Suppose we do a 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).**



- **One scenario is the following set of counts**

	fvc: abn	fvc: 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 in that "once smoking status is known, FVC status is known also"**

- **Another scenario is the following set of counts**

	fvc: abn	fvc: 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 we also observe that half (25) of the 50 *non*-smokers have an abnormal FVC, also.**
  - **This similarity in the data suggest that there is no association between smoking status and FVC**
  - **Put another way, the data suggest that lung function, as measured by FVC, is independent of smoking status.**

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

In Topic 7 (*Hypothesis Testing*), we used the idea of "proof by contradiction" to develop hypothesis tests. We do this in the contingency table setting, too.

Example, continued - Suppose we do a study to investigate the relationship between smoking and impairment of lung function, measured by forced vital capacity (FVC).

What are our null and alternative hypotheses?

- **Consider the notation that says**

  π_1 = the proportion of smokers with abnormal fvc
  π_2 = the proportion of non-smokers with abnormal fvc

- **The null hypothesis is that of independence, NO association, and says the proportion with abnormal fvc is the same, regardless of smoking status.**
  - $\pi_1 = \pi_2$

- **The alternative hypothesis is that of association/dependence and says the proportion with abnormal fvc will be a different number, depending on smoking status.**
  - $\pi_1 \neq \pi_2$

- **Thus,**

H_o: There is no association between the two variables, $\pi_1 = \pi_2$

"we'll argue proof-by-contradiction from here …"

H_a: The two variables are associated $\pi_1 \neq \pi_2$

"so as to advance this hypothesis"

Recall also from Topic 7 that a test statistic (also called *pivotal quantity*) is a comparison of what the data are to what we expected *under the assumption that the null hypothesis is correct* –

Introduction to observed versus expected counts.

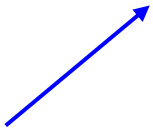
- **Observed counts are represented using the notation "O" or "n".**
- **Expected counts are represented using the notation "E"**

FVC

	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 subscripts -

- **The first subscript tells you the "row" (e.g. O_{21} is a cell count in row "2")**



- **The second subscript tells you "column" (e.g. O_{21} is a cell count in col "1")**

- **Thus, O_{21} is the count for the cell in row "2" and column "1"**

- **A subscript replaced with a "dot" is tells you what has been totaled over.**
  - **Thus, $O_{2.}$ is the total for row "2" taken over all the columns**
  - **Similarly $O_{.1}$ is the total for column "1", taken over all the rows**
  - **And $O_{..}$ is the grand total taken over all rows and over all columns**

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

	FVC: Abnormal	FVC: 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

What are the expected counts "E" under the assumption that the null hypothesis is true?

Hint – You already have an intuition for this. If a fair coin is tossed 20 times, the expected number of heads is 10. 10 represents (20 tosses) x (.50 probability of heads on each toss) = 10.

Solution:

- **Recall that we are utilizing the notation that says**

π_1 = the proportion of smokers with abnormal fvc

π_2 = the proportion of non-smokers with abnormal fvc

- **Under the assumption that the null of NO association/independence, then**
  - $\pi_1 = \pi_2 = \pi$ a common (*null hypothesis*) value

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

Hint – You already know this intuitively as = 45/100 = 0.45 = (total # with abnormal fvc)/(grand total).

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

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

- **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,**

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

- **So we expect 45% of the 50 smokers, or 22.5 persons, to have abnormal fvc, and we also expect 45% of the non-smokers, or 22.5 persons, to have abnormal fvc.** ***Yes, you are right. These expected counts are NOT whole integers. That's okay. Do NOT round.***

Expected # smokers w abnormal FVC = (#Smokers)($\hat{\pi}$) = (50)(.45) = 22.5 = E_{11}
Expected # NONsmokers w abnormal FVC = (#NONSmokers)($\hat{\pi}$) = (50)(.45) = 22.5 = E_{21}

- **We also need to obtain the expected counts of normal fvc.**
  - **We expect 55% of the 50 smokers, or 27.5, to have normal fvc, and we also expect 55% of the non-smokers, or 27.5, to have normal fvc.**
    Expected # smokers w normal FVC = (#Smokers)(1-$\hat{\pi}$) = (50)(.55) = 27.5 = E_{12}
    Expected # NONsmokers w normal FVC = (#NONSmokers)(1-$\hat{\pi}$) = (50)(.55) = 27.5 = E_{22}

- **Thus the following expected counts emerge.**

	FVC: Abnormal	FVC: 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

- E_{21}=22.5 E_{12}=27.5
- $E_{1.}$=50 $E_{.1}$=45

- **IMPORTANT NOTE about row totals and column totals -**
  - **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").

What is the test statistic (pivotal quantity)?

It is a chi square statistic. Examination reveals that it is defined to be a function of the comparisons of observed and expected counts. It can also be appreciated as a kind of "signal-to-noise" construction.

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

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

- **When the null hypothesis of no association is true, the observed and expected counts will be similar, their difference will be close to zero, resulting in a SMALL chi square statistic value.**
  - **Example – A null hypothesis says that a coin is "fair". Since a fair coin tossed 20 times is expected to yield 10 "heads", the result of 20 tosses is likely to be a number of observed heads that is close to 10.**
- **When the alternative hypothesis of an association is true the observed counts will be unlike the expected counts, their difference will be non zero and their squared difference will be positive, resulting in a LARGE POSITIVE chi square statistic value.**

  Example – The alternative hypothesis says that a coin is "NOT fair". An UNfair coin tossed 20 times is expected to yield a different number of heads that is NOT close to the null expectation value of 10.
- **Thus, evidence for the rejection of the null hypothesis is reflected in LARGE POSITIVE values of the chi square statistic.**

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

For reasons not detailed here (see Appendix), the comparison of observed and expected counts defined on page 9 is, often, distributed chi square *when the null is true*.

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

$$\frac{\left[\text{Observed Count} - \text{Expected Count}\right]^2}{\text{Expected Count}}$$

is distributed Chi Square (df = 1) approximately.

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

$$\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left[\text{Observed Count}(i,j) - \text{Expected Count}(i,j)\right]^2}{\text{Expected Count}(i,j)}$$

is distributed

Chi Square (df = [R-1][C-1]) approximately.

More on Degrees of Freedom *(What is the correct degrees of freedom?)*

Recall that "degrees of freedom" can be appreciated as the number of independent pieces of information. Here is how the idea works in contingency tables:

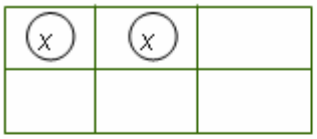
- **In a 2x2 table**

(x)	$n_1 - x$	n_1
$n_3 - x$	$n_2 - (n_3 - x)$	n_2
n_3	n_4	n

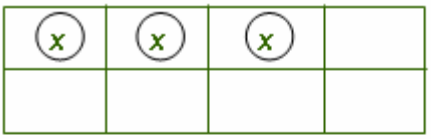
∴ we have "freedom" to fill in only one of the cells

⇒ 1 degree of freedom

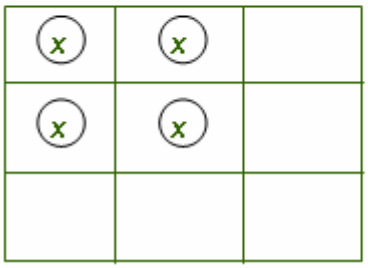
- **Now see what happens in larger tables**



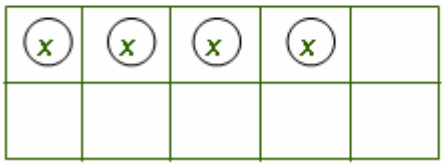
= 2 d.f.



= 3 d.f.



= 4 d.f.



= 4 d.f.

- **In each scenario, the last column is not free and the last row is not free.**

- **More generally,**

> **In an R x C table**
>
> **Degrees of freedom = (#rows – 1) x (#columns – 1)**
> **= (R – 1)(C – 1)**

We have the tools for computing the chi square test of association in a contingency table.

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.

	Last Consulted Physician			
Income	**≤ 6 months**	**7-12 months**	**>12 months**	**Total**
< $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**

The general layout uses either the "n_{ij}" or "O_{ij}" notation to represent the observed counts:

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

What are the π_{ij} now?

π_{ij} = the probability of being the combination that is income at level "i" and time since last consult at level "j"

Example: 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"

Example: $\pi_{1.}$ = probability [income is <$6000]

$\pi_{.j}$ = the *overall (marginal)* probability that time since last visit is at level "j"

Example: $\pi_{.1}$ = probability [time since last visit is $\leq$ 6 months]

What is independence now?

Again, you already have an intuition for this.
Recall the example of tossing a fair coin two times.
Because the outcomes of the two tosses are independent,
Probability of "heads" on toss 1 and "heads" on toss 2 = (.50)(.50) = .25

Now attach some notation to this intuition.

$\pi_{1.}$ = Probability of "heads" on toss 1
$\pi_{.2}$ = Probability of "heads" on toss 2
π_{12} = Probability of "heads" on toss 1 and "heads" on toss 2

Independence →
π_{12} = [probability heads on toss 1] x [probability heads on toss 2]
= [$\pi_{1.}$] [$\pi_{.2}$]

Thus, under independence

$$\pi_{ij} = [\ \pi_{i.}\ ]\ [\ \pi_{.j}\ ]$$

Pr["i" x "j" combination] = [Marginal "i" prob] x [Marginal "j"]

Under independence $\pi_{ij} = [\pi_{i.}][\pi_{.j}]$

Example, continued-

$\pi_{i.}$ = Probability that income is level "i"
$\pi_{.j}$ = 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,
$\pi_{ij} = [\pi_{i.}][\pi_{.j}]$

Assumptions of 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.**

Null and Alternative Hypotheses

$H_O: \pi_{ij} = \pi_{i.}\pi_{.j}$

$H_A: \pi_{ij} \neq \pi_{i.}\pi_{.j}$

Null Hypothesis Estimates of the π_{ij}

$\hat{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j}$ 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}

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

Test Statistic (Pivotal Quantity)

Just as we did before…

For <u>each cell</u>, departure of the observed counts from the null hypothesis expected counts is obtained using:

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

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

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

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

Under the null hypothesis

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

Decision Rule

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**

Computations

(1) For each cell, compute

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

(2) For each cell, compute

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

Example, continued -

Observed Counts (*this is just the table on page 12 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

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

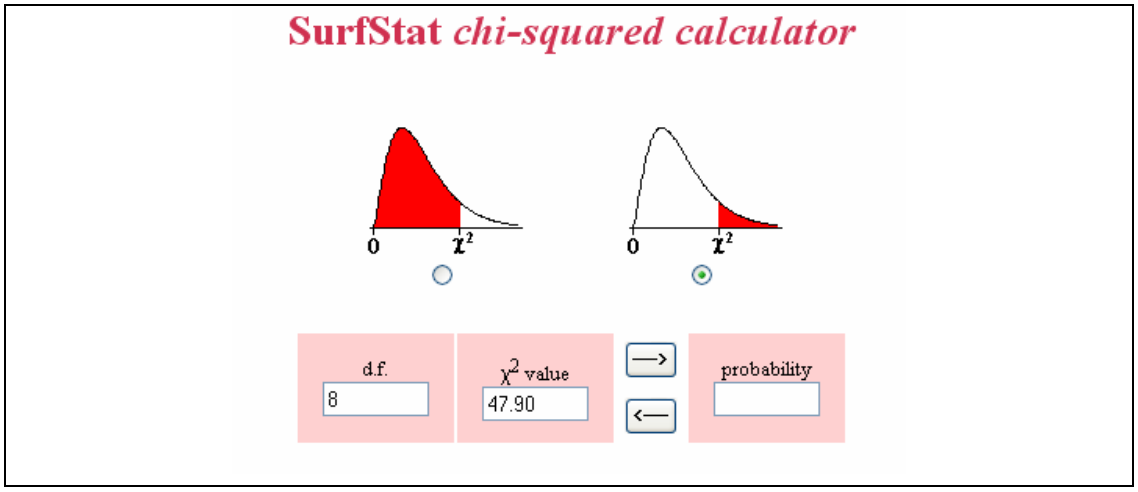
$$\chi^2_{(R-1)(C-1)} = \sum_{\text{all cells}} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] = \frac{(186-153.68)^2}{153.68} + \ldots + \frac{(259-241.22)^2}{241.22} = 47.90$$

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

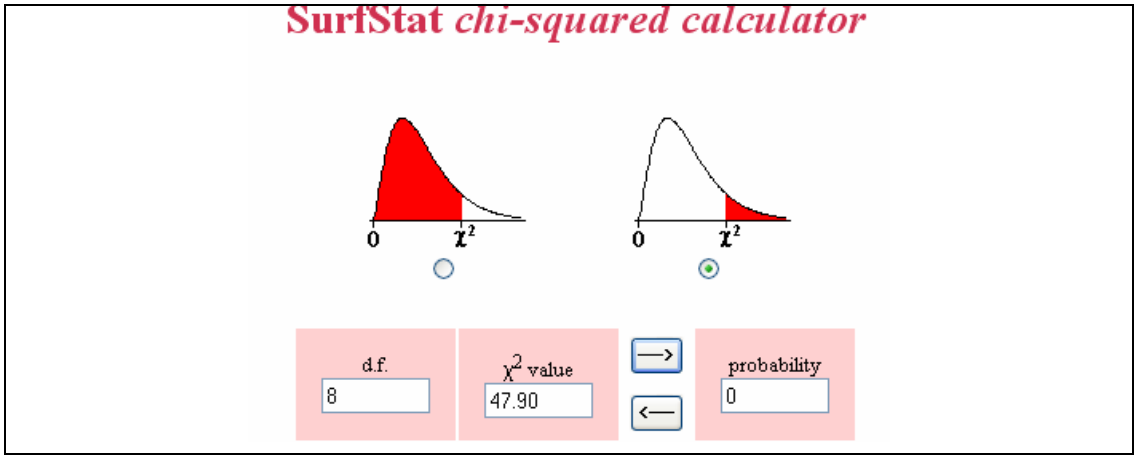
- **Statistical Decision Using Significance Level (p-value) Approach**
  http://www.anu.edu.au/nceph/surfstat/surfstat-home/tables/chi.php

  p-value = Probability [Chi square with df=8 ≥ 47.90] < < .0001
  Such an extremely small p-value is statistically significant →
  Reject the null hypothesis.

  To use: enter values in the boxes for "df" and chi square value". Click on RIGHT arrow"



You should then see "0" in the box labeled "probability"; thus, probability <<< .0001



Statistical Decision Using Critical Region Approach with type I error=.05

http://www.anu.edu.au/nceph/surfstat/surfstat-home/tables/chi.php

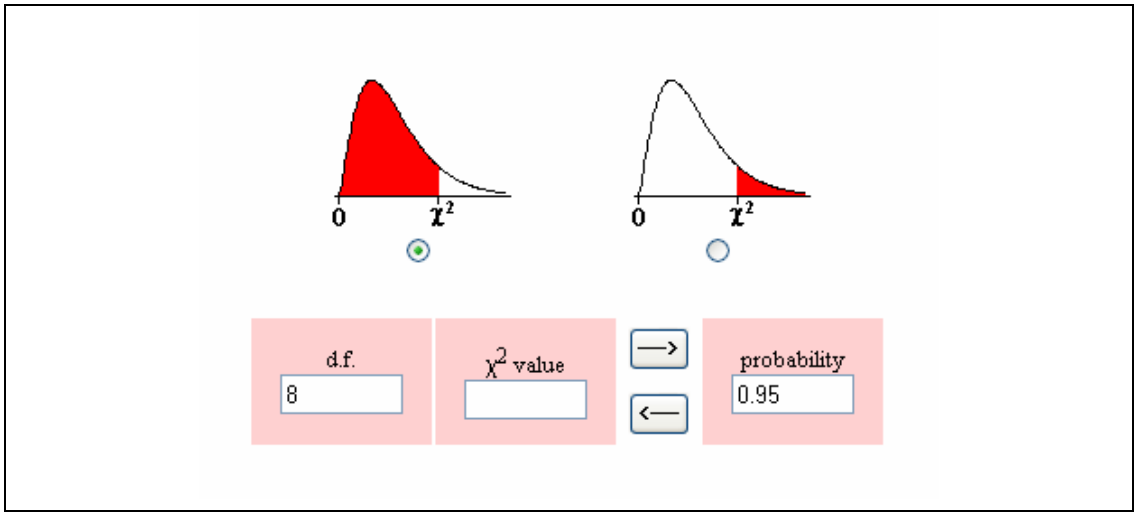
The critical region approach with type I error = .05 tells us to reject the null hypothesis for values of the test statistic that are greater than the 95th percentile. Equivalently, it tells us to reject the null hypothesis for values of the test statistic that are in the upper 5% of the distribution.

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

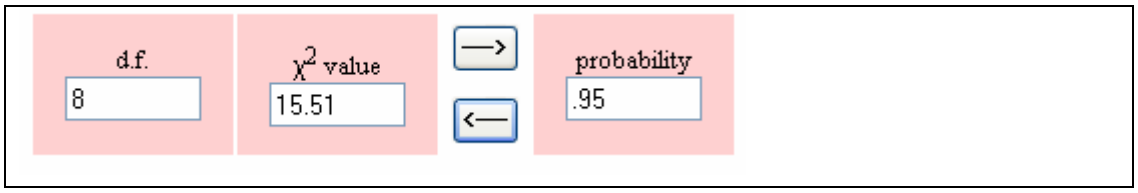
Observed statistic = 47.90 >> $\chi^2_{.95;df=8}$ = 15.51 →

Reject the null hypothesis.

To use: Click the radio dial button for "area to the left". Then enter values in the boxes for "df" and probability". Then click on LEFT arrow"



You should then see the value 15.51 in to box labeled chi square value



Example, continued -

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*.

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

Many epidemiology texts use a different notation for representing counts in the same chi square test of no association in a 2x2 table. 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 calculation for the chi square test that you've learned as being given by

$$\chi^2 = \sum_{\text{all cells}} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

is the same calculation as the following <u>shortcut</u> formula

$$\chi^2 = \frac{n\left(ad\text{-}bc\right)^2}{\left(a+c\right)\left(b+d\right)\left(c+d\right)\left(a+b\right)}$$

when the notation for the cell entries is the "a", "b", "c", and "d", above.

5. Hypotheses of Independence or No Association

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

For 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.

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 Topic 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 $Z\text{-score} = \frac{X-\mu}{\sigma}$	**{ Z-score }^{2}**	**1**
X_1, X_2, …, X_n are each distributed Normal (μ, σ^2) and are independent, so that $\overline{X}$ **is Normal (μ, σ^2/n) and** $Z\text{-score} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	**{ Z-score }^{2}**	**1**
X_1, X_2, …, X_n are each distributed Normal (μ, σ^2) and are independent and we calculate $S^2 = \frac{\sum_{i=1}^{n}(X-\overline{X})^2}{n-1}$	$\frac{(n-1)S^2}{\sigma^2}$	**(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.

	$Z_1, Z_2, \ldots, Z_n$ are each Bernoulli with probability of event = π. $$E[Z_i] = \mu = \pi$$ $$Var[Z_i] = \sigma^2 = \pi(1-\pi)$$ $\downarrow$	
	1. The net number of events $X = \sum_{i=1}^{n} Z_i$ is Binomial (N,π) **2. We learned previously that the distribution of the average of the Z_i is well described as Normal(μ, σ^2/n).** **Apply this notion here: By convention,** $$\overline{Z} = \frac{\sum_{i=1}^{n} Z_i}{n} = \frac{X}{n} = \overline{X}$$ $\downarrow$	
	3. So perhaps the distribution of the sum is also well described as Normal. At least approximately **If $\overline{X}$ is described well as Normal (μ, σ^2/n)** **Then $X = n\overline{X}$ is described well as Normal ($n\mu$, $n\sigma^2$)** $\downarrow$	
	Exactly: X is distributed Binomial(n,π) **Approximately: X is distributed Normal ($n\mu$, $n\sigma^2$)** **Where: $\mu=\pi$ and $\sigma^2 = \pi(1-\pi)$**	

Putting it all together …

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

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

- **Preliminaries**

  X = "Observed" = O
  nπ = "Expected" = E

- **As n gets larger and larger**

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

- **Upon substitution,**

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

Thus,

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

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

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**

$$\sum_{i=1}^{R}\sum_{J=1}^{C}\frac{\left[\begin{array}{ccc}\text{Observed} & & \text{Expected}\\ \text{Count}_{ij} & - & \text{Count}_{ij}\end{array}\right]^2}{\text{Expected Count}_{ij}}$$

is Chi Square [df =(R-1)(C- 1)] approx.