

Unit 11
Chi Square Tests
plus Fisher's Exact Test

"I shall never believe that God plays dice with the world"

- Albert Einstein (1879-1955)

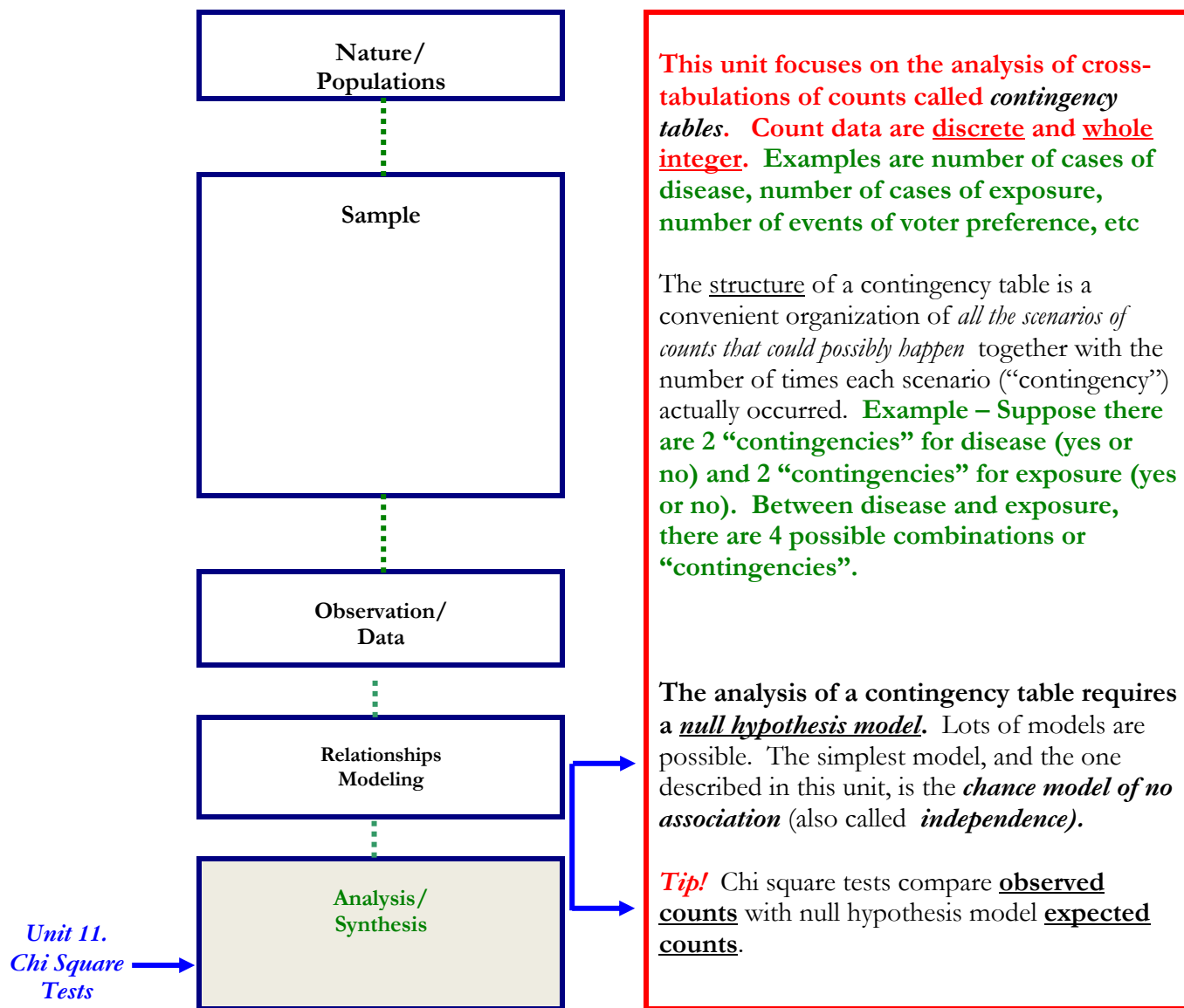
How many patients died? How many travelers on a cruise ship were exposed to contaminated water? This unit is about **counts**.

This unit addresses such questions as: **Are there too many (or too few) events compared to what I might have expected by chance?**

Table of Contents

Topic		
1. Unit Roadmap		3
2. Learning Objectives		4
3. Introduction to Contingency Tables		5
a. Contingency Table Counts and Notation		5
b. Contingency Table Degrees of Freedom		8
4. Null Hypothesis of Independence or No Association		10
5. Tests of No Association for a 2x2 Table		11
a. Chi Square Test		12
b. Fisher's Exact Test		18
6. (For Epidemiologists) Special Case: More on the 2x2 Table		22
7. The Chi Square Test of No Association in an R x C Table		23
Appendices		
A1. Relationship Between the Normal(0,1) and the Chi Square Distribution ...		33
A2. Fisher's Exact Test – Derivation		37

1. Unit Roadmap



2. Learning Objectives

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

- Identify settings where the chi square test of no association 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 null hypothesis “expected” counts;
- Calculate, by hand, the chi square test of “no association” for a 2x2 table of observed frequencies ;
- Perform a Fisher's Exact Test of “no association” for a 2x2 table where the cell frequencies are small;
- 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

In a contingency table analysis, we compare observed numbers of events (a count) with some null hypothesis expected numbers of events.

- **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. Is this statistically significant?
- **Example - Is high income associated with membership in the Republican party?**
The number of wealthy persons (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. Is this statistically significant?

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 of smoking status and FVC status. **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

Fixed by sample size

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 50 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.*

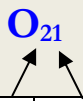
Introduction to observed versus expected counts.

- 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”

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

How to read subscripts that are dots-

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

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

- **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

Example, continued – In this sample of 100 ($O_{..} = 100$), there are 45 with an abnormal FVC (column 1 total is $O_{.1} = 45$), 50 smokers (row 1 total is $O_{1.} = 50$). There are 40 who are smokers with an abnormal FVC ($O_{11} = 40$). And so on.

In the next section, we'll learn about the null hypothesis 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

$$= (\text{\#rows} - 1) * (\text{\#columns} - 1)$$

$$= (R - 1) * (C - 1)$$

4. Null Hypothesis of Independence or No Association

“Independence”, “No Association”, “Homogeneity of Proportions” are various wordings of the same null hypothesis.

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.

5. Tests of No Association for a 2x2 Table

Example -

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	$O_{21}=5$	$O_{22}=45$	$O_{2.}=50$	
	$O_{.1}=45$	$O_{.2}=55$	$O_{..}=100$	

Among the 50 smokers, a disproportionately large number, 40 (80%), have an abnormal FVC. By comparison, among the 50 who don't smoke, there are just 5 instances of abnormal FVC (10%) among the non-smokers. Do these data provide statistically significant evidence of an association of smoking with abnormal FVC?

Recall from Unit 8 (*Statistical Literacy – Estimation and 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)

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

5a. Chi Square Test

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. Under the null hypothesis, it is distributed Chi Square.

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

(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 expected counts.**

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

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

(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 = 27.5 persons, to have normal fvc; and
55% of the 50 non-smokers = 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		$E_{21}=22.5$	$E_{22}=27.5$	$E_{2.}=50$
		$E_{.1}=45$	$E_{.2}=55$	$E_{..}=100$

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

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”).

An Easy Solution for the Null Hypothesis Expected Counts
H₀: “No Association”

Expected Count in row “i” and column “j”

$$= E_{ij} = \frac{(O_{i.})x(O_{.j})}{O_{..}}$$

$$= \frac{(\text{row “i” total})x(\text{column “j” total})}{\text{Grand total}}$$

Example: $E_{12} = \frac{(\text{Row “1” total})(\text{Column “2” total})}{\text{Grand total}} = \frac{(50)(55)}{100} = 27.5$

Provided the cell frequencies sizes are sufficiently large (see suggested "rule" below), the appropriate test statistic is a chi square statistic defined as follows

The chi square statistic here is a comparison of observed and null hypothesis expected counts.

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

Suggested "Rule":

It is okay to Perform a Chi Square Test for a 2x2 Table if:

ALL of the null hypothesis **expected frequencies** (E_{ij}) are greater than 5

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.

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

$$\begin{aligned}\text{Chi Square}_{df=1} = \chi^2_{df=1} &= \sum_{\text{all cells "i,j"}} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] \\ &= \sum_{\text{all cells "i,j"}} \left[\frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}} \right]\end{aligned}$$

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

Perform the Calculations.

Recall the observed and null hypothesis expected counts.

Observed Counts, "O"

		FVC		
		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"

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

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

$$= 49.4949$$

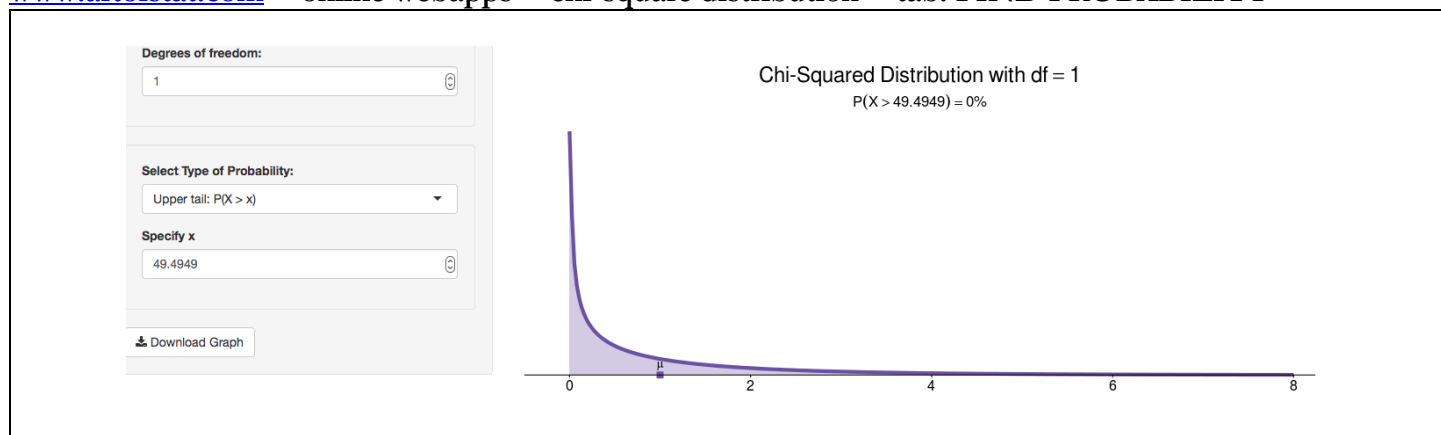
P-Value Calculation

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

<<<<.0001

Note – This calculator returns probability=0 here

www.artofstat.com > online webapps > chi-square distribution > tab: FIND PROBABILITY



<https://istats.shinyapps.io/ChisqDist/>

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

R Users

Input 2x2 Table

```
table_page11 <- as.table(rbind(c(40,10),c(5,45))) # key: rbind( c(a=40,b=10), c(c=5,d=45)
dimnames(table_page11) <- list(
  SMOKE=c("Smoke","Dont Smoke"),
  FVC=c("Abnormal","Normal"))
table_page11

##           FVC
## SMOKE      Abnormal Normal
## Smoke           40      10
## Dont Smoke       5      45
```

Chi Square Test of 2x2 Table WITHOUT Continuity Correction

```
options(scipen=1000)
chisq.test(table_page11,correct=FALSE)

##
## Pearson's Chi-squared test
##
## data:  table_page11
## X-squared = 49.495, df = 1, p-value = 0.000000000001989
```

Assumption of the null hypothesis has led to an unlikely result. Reject the null.

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.

The Chi Square Test is Appropriate for Moderate to Large Sample Size Tables Only.
For small sample size tables, use the Fisher's Exact Test instead.

Different texts and sources suggest different “rules of thumb”. They’re similar. Here is a suggested guideline:

Perform a Fisher's Exact Test for a 2x2 Table if:

One or more of the null hypothesis expected frequencies (E_{ij}) is 5 or less.

It is okay to Perform a Chi Square Test for a 2x2 Table if:

ALL of the null hypothesis expected frequencies (E_{ij}) are greater than 5

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

5b. Fisher's Exact Test

Perform a Fisher's Exact Test for a 2x2 Table if:
One or more of the null hypothesis **expected frequencies (E_{ij})** is **5 or less**.

Fisher's exact test for a 2x2 table tests the same hypothesis as that tested by the Chi Square Test of no association for a 2x2 table. Perform a Fisher's Exact Test for small sample size 2x2 tables; eg – when one or more of the expected cell frequencies is less than 5.

Same Example (see again page 11):

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

		<u>FVC</u>		
		Abnormal	Normal	
Smoke	40	10	50	
Don't smoke	5	45	50	
	45	55	100	

Among the 50 smokers, a disproportionately large number, 40 (80%), have an abnormal FVC. By comparison, among the 50 who don't smoke, there are just 5 instances of abnormal FVC (10%) among the non-smokers. Do these data provide statistically significant evidence of an association of smoking with abnormal FVC?

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.

$$H_0: \pi_1 = \pi_2$$

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

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. The Fisher's Exact test null hypothesis model is the Central Hypergeometric Distribution.

The probability model underlying the Fisher Exact Test is presented in more detail in BIOSTATS 640 course notes, **2. Discrete Distributions**.

In brief, the null hypothesis probability model that underlies Fisher's Exact test treats the row and column totals as fixed. Because of this, only one cell count is free to vary. The remaining cell counts are then obtained by subtraction from their corresponding row and column totals.

The Fisher Exact Test for a 2x2 Table Uses the "a, b, c, d" notation for the cell counts.

The counts are "a", "b", "c", and "d" as follows. By convention, we treat the row 1 column 1 cell count "a" as the random variable. All the rest are obtained by subtraction from the fixed row and column totals.

		<u>Column Variable</u>		
		yes	no	
<u>Row Variable</u>	yes	a	b	(a+b)
	no	c	d	(c+d)
		(a+c)	(b+d)	N

Under the **null hypothesis of "no association"**, the probability of obtaining any particular set of counts "a", "b", "c", and "d" subject to the constraints that the row totals and the column totals (and hence the grand total) is a **central hypergeometric probability distribution** calculation. For now, we'll let the computer do this for us.

5. How to solve for the Fisher Exact Test p-value.

Note!!! The calculations of a Fisher Exact test do NOT yield a test statistic value. They yield only a p-value result.

The idea of this calculation is the same as that for p-value calculations in the testing settings that you have already seen. Here it takes the following form:

$$\text{p-value} = \Pr [\text{count in row 1 column 1} = a \text{ or more extreme} \mid \text{null is true}]$$

Example, continued -

In this example, because the column "1" total is 45, the more extreme values for "a" are 41, 42, 43, 44, and 45.

$$\begin{aligned} \text{p-value} = & \Pr[\text{table with } a=40] + \Pr[\text{table with } a=41] + \Pr[\text{table with } a=42] \\ & + \Pr[\text{table with } a=43] + \Pr[\text{table with } a=44] + \Pr[\text{table with } a=45] \end{aligned}$$

For the interested reader, the details are shown in Appendix A2.

We'll let the computer do these calculations for us!

www.artofstat.com > online webapps > Fisher's Exact Test

You will be given a blank 2x2 table. Enter your counts .

Fisher's Exact Test

Enter counts for 2 x 2 table:

40	10
5	45

Select alternative hypothesis:

odds ratio not equal to 1 (association)

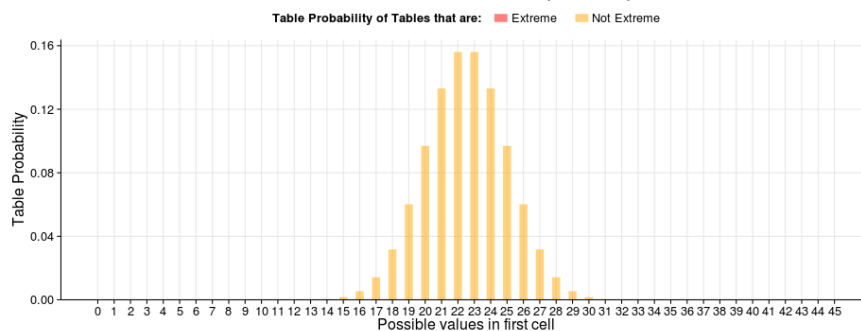
[Download Graph](#)

Click **RUN TEST**

You should see the following

Test	Value of Test Statistic	Alternative Hypothesis	P-value
Fisher's Exact Test	40	odds ratio not equal to 1 (association)	0.0000

Distribution of Test Statistic (First Cell)



Wow. The probability of a count in the “first cell” (row 1, column 1) ≥ 40 is so small, we can't even see red.

R Users

Input 2x2 Table

```
table_page11 <- as.table(rbind(c(40,10),c(5,45)))
dimnames(table_page11) <- list(
  SMOKE=c("Smoke","Dont Smoke"),
  FVC=c("Abnormal","Normal"))
table_page11
##           FVC
## SMOKE      Abnormal Normal
## Smoke           40      10
## Dont Smoke       5      45
```

Fisher Exact Test of 2x2 Table w ONE sided alternative (Positive association)

```
options(scipen=1000)
fisher.test(table_page11,alternative="greater")
##
## Fisher's Exact Test for Count Data
##
## data:  table_page11
## p-value = 0.0000000000003638
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
## 11.95 Inf
## sample estimates:
## odds ratio
## 34.04563
```

6. Evaluate Findings and Report.

Under the null hypothesis assumption of no association of smoking with abnormal forced vital capacity, the Fisher's Exact Test p-value is $p\text{-value} = \Pr [a \geq 40]$ is less than 0.0001. 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.** Conclude that this sample provides statistically significant evidence of an association of smoking with abnormal forced vital capacity.

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.

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

The “O” and “E” formula for the Chi Square 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 Chi Square 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. 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,

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

- Summed over all cells in an R x C table, when the null 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

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

$$\text{Degrees of Freedom} = DF = (R-1)(C-1)$$

$$\text{where, Expected Count(i,j)} = \frac{(\text{row "i" total})(\text{column "j" total})}{\text{Grand total}}$$

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.

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

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 ≤ 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 ≤ 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

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

Let

$\pi_{1.}$ = Probability of “heads” on toss 1, regardless of outcome of toss 2

$\pi_{.2}$ = Probability of “heads” on toss 2, regardless of outcome on toss 2

Now let

π_{12} = Probability of “heads” on toss 1 and “heads” on toss 2

Independence \rightarrow

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

Thus, under the null hypothesis of independence

$$\begin{array}{ccccc} \pi_{ij} & = & [\pi_{i.}] & [\pi_{.j}] \\ \downarrow & & \downarrow & \searrow & \\ \text{Pr}[i \times j \text{ combination}] & = & [\text{Marginal “i” probability}] & \times & [\text{Marginal “j” probability}] \end{array}$$

Application of the null hypothesis of independence to the RxC Setting: For example

Let

$\pi_{1.}$ = Probability that income is $< \$6000$, overall

$\pi_{.1}$ = Probability that consult time is ≤ 6 months, overall

Now let

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

Null hypothesis of independence says that \rightarrow

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

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

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,

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

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,

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

Behavior of the Test Statistic when the null hypothesis is **NOT** true

When the null hypothesis is **not** true, at least some of the differences (observed – expected) will be very different from zero. In this scenario, application of the null hypothesis model to the actual data will lead to an unlikely result, namely:

- The chi square statistic value will be LARGE; **and**
- The p-value calculation, using the null model, will be a SMALL probability; **and**
- The observed chi square test statistic will excel the CRITICAL VALUE threshold.

Perform the Calculations Using the Null Hypothesis Model of Independence

(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

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

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] $<$.0001

A (p-value) probability of less than 1 chance in 10,000 is a very unlikely event and quite a challenge to the assumption of the null hypothesis! Therefore, we will say that it is statistically significant and *reject the null hypothesis*.

Again, we'll let the computer do these calculations for us ...

www.artofstat.com > online webapps > Chi Square Test > Tab: Test of Independence/Homogeneity

Preliminary: First specify number of rows=5 and number of columns=3

You will then be given a blank 5x3 table. Enter your counts .

The Chi-Squared Test

Select dimension of contingency table:

Number of Rows (at least 2):

Number of Columns (at least 2):

Enter counts for table:

186	38	35
227	54	45
219	78	78
355	112	140
653	285	259

Submit Counts

Test of Independence/Homogeneity

Goodness of Fit

Click **Submit Counts**

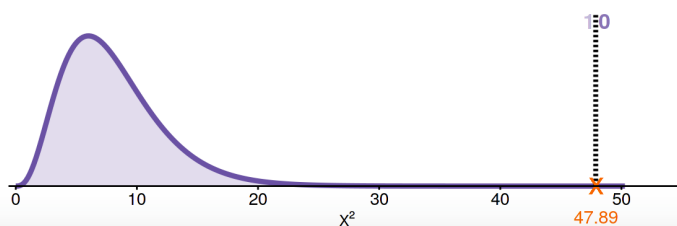
I've omitted some of what you'll see. Towards the bottom you should see the following

Pearson's Chi Squared Test of Independence/Homogeneity:

Null Hypothesis	Alternative Hypothesis	Test Statistic χ^2	df	P-value
No Association	Association	47.8923	8	0.0000

Chi-Squared Distribution with df = 8

$\chi^2 = 47.9$, df = 8, P-value = 0



Note – P-value=0 allows us to say p-value <<<< .0001

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

R Users

Input RxC Table

```
table_page24 <- as.table(rbind(c(186,38,35),c(227,54,45),c(219,78,78),c(355,112,140),c(653,285,259)))
dimnames(table_page24) <- list(
  INCOME=c("<$6000","$6000-$9999","$10000-$13999","$14000-$19999", "> $20000"),
  LAST_CONSULTED_PHYSICIAN=c("<= 6 mos","7-12 mos", "> 12 mos"))
table_page26
```

```
##          LAST_CONSULTED_PHYSICIAN
## INCOME    <= 6 mos 7-12 mos > 12 mos
## <$6000      186    38    35
## $6000-$9999 227    54    45
## $10000-$13999 219    78    78
## $14000-$19999 355   112   140
## > $20000    653   285   259
```

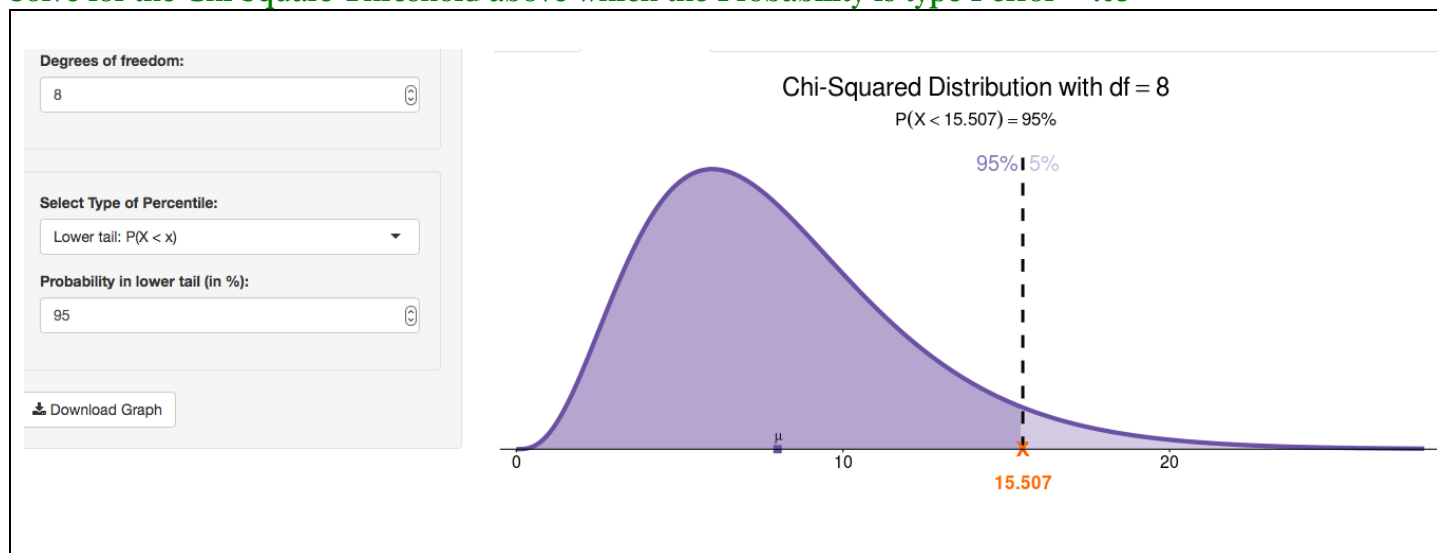
Chi Square Test of RxC Table (NOTE: there is no continuity correction for tables with >2 rows, > 2 col)

```
options(scipen=1000)
chisq.test(table_page24)
```

```
##
## Pearson's Chi-squared test
##
## data:  table_page26
## X-squared = 47.892, df = 8, p-value = 0.0000001036      Reject the null.
```

Chi Square Test of No Association Using a Critical Region Approach with (type 1 error = 0.05)

Solve for the Chi Square Threshold above which the Probability is type I error = .05



<https://istats.shinyapps.io/ChisqDist/>

$\chi^2_{.95;df=8} = 15.507$ is our critical value.

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

Compare the Observed Chi Square Statistic to this Threshold. Is it Larger?

The observed statistic = **47.90** obtained on the previous page is **larger than** $\chi^2_{.95;df=8} = 15.507$. Thus, it falls in the critical region of “unlikely under the null hypothesis model”

→ Statistical rejection of 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. → ***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 BIOSTATS 640, *Intermediate Biostatistics*.

Appendix A1.
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 8, *Statistical Literacy – Estimation and Hypothesis Testing*), 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 $Z\text{-score} = \frac{X - \mu}{\sigma}$	$\{ Z\text{-score} \}^2$	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 $Z\text{-score} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$\{ Z\text{-score} \}^2$	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>If \bar{X} is described well as Normal ($\mu, \sigma^2/n$)</p> <p>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>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
X has a distribution that is <u>Binomial</u> (n, π) <u>exactly</u>	<p>X has a distribution that is <u>Normal</u> ($n\mu, n\sigma^2$) <u>approximately</u>, where</p> $\mu = \pi$ $\sigma^2 = \pi(1-\pi)$ <p>↓</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>↓</p>	
	{ 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 $\{Z\text{-score}\}^2$ 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.

Appendix A2.
Fisher's Exact Test – Derivation
For the interested reader

The probability model underlying the Fisher Exact Test is presented in more detail in BIOSTATS 640, *Intermediate Biostatistics*.

In brief, the null hypothesis probability model that underlies Fisher's Exact test treats the row and column totals as fixed. Because of this, only one cell count is free to vary. The remaining cell counts are then obtained by subtraction from their corresponding row and column totals.

The Fisher Exact Test for a 2x2 Table Uses the “a, b, c, d” notation for the cell counts.

The counts are “a”, “b”, “c”, and “d” as follows. By convention, we treat the row 1 column 1 cell count “a” as the random variable.

		<u>Column Variable</u>		
		yes	no	
<u>Row Variable</u>	yes	a	b	(a+b)
	no	c	d	(c+d)
		(a+c)	(b+d)	N

Under the null hypothesis of “no association”, the probability of obtaining any particular set of counts “a”, “b”, “c”, and “d” subject to the constraints that the row totals and the column totals (and hence the grand total) are fixed is

$$\text{Pr}[\text{arrangement of "a","b","c","d"}|\text{null hypothesis model}] = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!}$$

Where the “*exclamation*” notation is the factorial notation and has the following meaning:

- a! reads “a factorial”
- a! = product of a and all the whole numbers less than it, down to unity. Thus,
- a! = a x (a-1) x (a-2) x x (2) x 1
- By convention, 0! = 1

Example -

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

How to solve for the Fisher Exact Test p-value.

The Fisher Exact Test p-value is the sum of the probabilities using the formula on the previous page, taken over the observed table plus all the other configurations of frequencies (“a”, “b”, “c”, and “d”) that give as much or more evidence of an association, each time keeping the row and column totals fixed.

2 x 2 Table
Fisher Exact Test of No Association
 Use when: 1 or more null hypothesis expected counts is ≤ 5

$$p\text{-value} = \sum_{\text{Tables with same or greater evidence of association}} \Pr["a", "b", "c", "d" | \text{null}] = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!}$$

where $a! = a \times (a-1) \times (a-2) \times \dots \times (2) \times (1)$

Rejection of the null hypothesis of “no association” occurs for small values of p-value

Perform the Calculations.

Null hypothesis model probability of the **observed table; with a=40:**

	FVC		
	Abnormal	Normal	
Smoke	40	10	50
Don't smoke	5	45	50
	45	55	100

$$\Pr[\text{observed, with } a=40] = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!} = \frac{(50)! (50)! (45)! (55)!}{40! 10! 5! 45! 100!}$$

$$= 3.542 \times 10^{-13}$$

Null hypothesis model probability of **the more extreme table; with a=41:**

	FVC		
	Abnormal	Normal	
Smoke	41	9	50
Don't smoke	4	46	50
	45	55	100

$$\Pr[\text{table with } a=41] = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!} = \frac{(50)! (50)! (45)! (55)!}{41! 9! 4! 46! 100!}$$

$$= 9.390 \times 10^{-15}$$

Null hypothesis model probability of the more extreme table; with a=42:

		<u>FVC</u>		
		Abnormal	Normal	
Smoke	42	8	50	
Don't smoke	3	47	50	
		45	55	100

$$\Pr[\text{table with } a=42] = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!} = \frac{(50)! (50)! (45)! (55)!}{42! 8! 3! 47! 100!}$$

$$= 1.712 \times 10^{-16}$$

Null hypothesis model probability of the more extreme table; with a=43:

		<u>FVC</u>		
		Abnormal	Normal	
Smoke	43	7		50
Don't smoke	2	48		50
		45	55	100

$$\Pr[\text{table with } a=43] = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!} = \frac{(50)! (50)! (45)! (55)!}{43! 7! 2! 48! 100!}$$

$$= 1.991 \times 10^{-18}$$

Null hypothesis model probability of the more extreme table; with a=44:

		<u>FVC</u>		
		Abnormal	Normal	
Smoke	44	6	50	
Don't smoke	1	49	50	
		45	55	100

$$\Pr[\text{table with } a=44] = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!} = \frac{(50)! (50)! (45)! (55)!}{44! 6! 1! 49! 100!}$$

$$= 1.293 \times 10^{-20}$$

Null hypothesis model probability of the more extreme table; with a=45:

		<u>FVC</u>		
		Abnormal	Normal	
Smoke	45	5	50	
Don't smoke	0	50	50	
		45	55	100

$$\Pr[\text{table with } a=45] = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!} = \frac{(50)! (50)! (45)! (55)!}{45! 5! 0! 50! 100!}$$

$$= 3.448 \times 10^{-23}$$

$$\begin{aligned} \text{p-value} = & \Pr[\text{table with } a=40] \\ & + \Pr[\text{table with } a=41] \\ & + \Pr[\text{table with } a=42] \\ & + \Pr[\text{table with } a=43] \\ & + \Pr[\text{table with } a=44] \\ & + \Pr[\text{table with } a=45] \end{aligned}$$

$$\begin{aligned} = & 3.542 \times 10^{-13} \\ & + 9.390 \times 10^{-15} \\ & + 1.712 \times 10^{-16} \\ & + 1.991 \times 10^{-18} \\ & + 1.293 \times 10^{-20} \\ & + 3.448 \times 10^{-23} \end{aligned}$$

<<<< .0001 Not surprising; this matches the p-value for the chi square test on page 16.