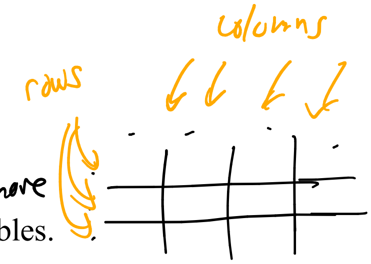
Chapter 12: Inference on Categorical Data

Section 12.2: Tests for Independence and the ~~Homogeneity of Proportions~~



CONTINGENCY TABLES

Def A **contingency table** is a table in which frequencies correspond to two or more variables.

One variable is used to categorize rows, and a second variable is used to categorize columns.

TEST FOR INDEPENDENCE

Def A **test of independence** tests the null hypothesis that in a contingency table, the row and column variables are independent.

Steps for Hypothesis Test for Independence

What to Find…

- Number of categories, k
- Expected Counts, E_i

$$E = (column\ total) \cdot \left(\frac{row\ total}{table\ total}\right)$$

Requirements:

- The sample data are randomly selected. SRS
- The sample data are represented as frequency counts in a two-way table.
- For each cell in the contingency table, the expected frequency is at least 5. $E_i \geq 5$

Step 1: *Hypotheses*

H_0: the row and column variables are independent (no association)

H_A: the row and column variables are dependent (some association)

ALWAYS RIGHT **-TAILED TEST!**

Step 2: *Level of Significance*

α

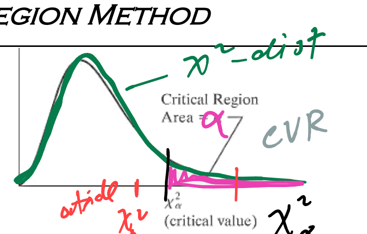
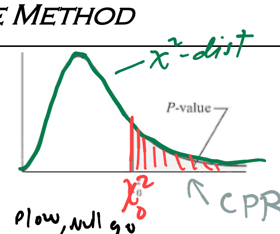
$\alpha = 0.05$ if not given.

Step 3: *Test Statistic*

$$\chi_0^2 = \sum \frac{(O-E)^2}{E} \quad \text{and} \quad df = (\underset{\#\text{ of rows}}{r} - 1) \cdot (\underset{\#\text{ columns}}{c} - 1)$$

Step 4: *Find a Critical Value or P-Value to check using either the Critical Value or P-value method.*

Step 5: *Make a decision and draw a conclusion.*

CRITICAL REGION METHOD	P-VALUE METHOD
* Table VIII $df = (r-1)(c-1)$  $\left\{\begin{array}{l} \textit{Reject } H_0 \text{ if } \chi^{2*} \text{ lies in the critical region} \\ \textit{Fail to Reject } H_0 \text{ if } \chi^{2*} \text{ doesn't lie in the critical region} \end{array}\right.$ (inside)	$df = (r-1)(c-1)$  $\left\{\begin{array}{l} \textit{Reject } H_0 \text{ if } P\text{-value} \leq \alpha \\ \textit{Fail to Reject } H_0 \text{ if } P\text{-value} > \alpha \end{array}\right.$ P low, null go / P high, null fly

GRAPHING CALCULATOR (TI-83 OR 84)

χ^2cdf(low, high, df) (New!)

Instructions: (a) 2nd $\Rightarrow x^{-1}$(matrix)$\Rightarrow$EDIT$\Rightarrow$1: $[A] \Rightarrow$ENTER

(b) Input # of rows × # of columns ($r \times c$)

(c) Enter the sample values

(d) STAT $\Rightarrow$ TESTS $\Rightarrow$ C: χ^2-Test (New!)

Ex 1: A pharmaceutical company claims that they have created to two different pills that help with the flu epidemic. Adult Americans are randomly assigned to three different groups: one group takes pill 1, one group takes pill 2 and the last group takes a sugar pill. It was then observed whether or not each person got the flu or not. Below is a contingency table with the observations made:

[A]

#1 row total

Observed (O)	*Pill 1*	*Pill 2*	*Placebo*	
Flu	20 O	30	30	$\Sigma =$ 80
Not Flu	100	110	90	$\Sigma =$ 300
	$\Sigma =$ 120	$\Sigma =$ 140	$\Sigma =$ 120	$\Sigma =$ 380

column #1 total?

table total

We need to come up with an expected frequency count.

(a) What is the probability of choosing a person who has the flu?

$P(flu) = \frac{80}{380} = 0.211$

(b) What is the probability of choosing a person who does not have the flu?

$P(flu^c) = P(\text{no flu}) = 1 - \frac{80}{380} = 1 - 0.211 = 0.789$

(c) Of the people who took pill 1, how many would you *expect* to have the flu?

$flu \mid Pill 1 = 120 * 0.211 = 25.32$

26 people

(d) Of the people who took pill 1, how many would you *expect* to not have the flu?

$flu^c \mid Pill 1 = 120 - 26 = 94$

$E = (\text{column total}) \cdot \left(\frac{\text{row total}}{\text{table total}}\right)$

Expected Table (E)	*Pill 1*	*Pill 2*	*Placebo*
Flu	$120\left(\frac{80}{380}\right) = 25.26$	$140\left(\frac{80}{380}\right) = 29.47$	$120\left(\frac{80}{380}\right) = 25.26$
Not Flu	$120\left(\frac{300}{380}\right) = 94.74$	$140\left(\frac{300}{380}\right) = 110.53$	$120\left(\frac{300}{380}\right) = 94.74$

HT

Test the claim that the type of pill is independent of getting the flu. Let $\alpha = 0.1$

Check requirements 1) SRS ✓ 2) 2way table ✓ 3) $E_i > 5$ ✓

Null and Alternative Hypothesis

H_0 : the rows & columns are independent.

H_A : the rows & columns are dependent.

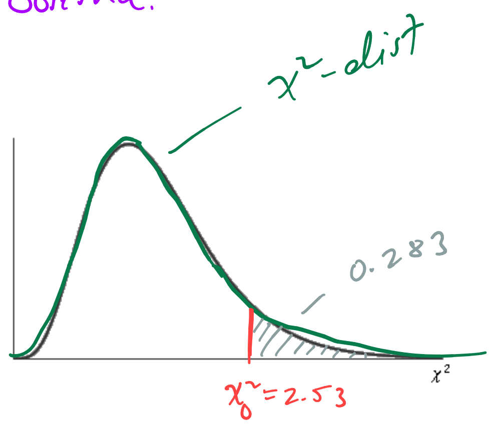
Test Statistic Calculator! χ^2-Test & matrix [A] = Observed.

$\boxed{\chi_0^2 = 2.53}$

P-Value/Critical Region:

Calculator: $\boxed{P = 0.283}$

$df = (r-1)\cdot(c-1)$
$= (2-1)\cdot(3-1)$
$= 1 \cdot 2 = 2$



Decision about Null Hypothesis

$\alpha = 0.1$

$P = 0.283$

$P > \alpha \longrightarrow$ P high, Null fly

$\boxed{\text{Fail to Reject } H_0}$

Conclusion

"There is not enough statistical evidence to support the claim that the type of pill is dependent on getting the flu."

Ex 2: Many people believe that criminals who plead guilty get lighter sentences than those who are convicted in trials. The accompanying table summarizes randomly selected sample data for San Francisco defendants in burglary cases. All of the subjects had prior prison sentences.

O	*Guilty Plea*	*Not Guilty Plea*	
Sent to Prison	390	60	$\Sigma = 450$
Not Sent to Prison	561	17	$\Sigma = 578$
	$\Sigma = 951$	$\Sigma = 77$	$\Sigma = 1028$

α

At the 0.05 significance level, test the claim that the sentence is independent of the plea.

HT

Find the Expected Frequencies

$E = \text{column total}\left(\frac{\text{row total}}{\text{table total}}\right)$

E	*Guilty Plea*	*Not Guilty Plea*
Sent to Prison	$951\left(\frac{450}{1028}\right) = 416.29$	$77\left(\frac{450}{1028}\right) = 33.71$
Not Sent to Prison	$951\left(\frac{578}{1028}\right) = 534.71$	$77\left(\frac{578}{1028}\right) = 43.29$

Check requirements 1) SRS ✓ 2) 2way table ✓ 3) $E_i \geq 5$ ✓

Null and Alternative Hypothesis

H_0: row & column variables are independent.

H_A: row & column variables are dependent.

Test Statistic

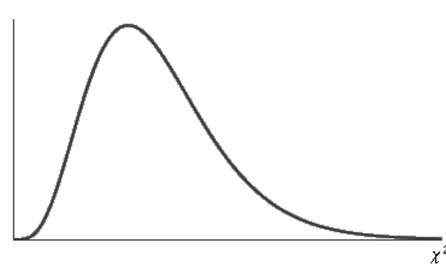
Calculator

- [A] "observed"
- χ^2-Test

$\boxed{\chi_0^2 = 39.43}$

P-Value

$\boxed{P = 3.39 \times 10^{-10}} = 0.000\,000\,000\,339.$ 0+



Decision about Null Hypothesis

$\alpha = 0.05$

$P = 3.39 \times 10^{-10}$

$P < \alpha \rightarrow$ P low, Null go!

$\boxed{\text{Reject } H_0}$

Conclusion

"There is enough statistical evidence to support the claim that the sentence is dependent on the plea."

TEST FOR HOMOGENEITY OF PROPORTIONS (ONLY IF TIME PERMITS) (SKIP)

Def A **test of homogeneity of proportions** tests whether different populations have the same proportion of individuals with some characteristic.

In simple terms, this means we run a hypothesis test for two (or more) proportions from samples that may not be independent (in other words, just from many categories).

For example, we might consider the proportion of individuals who get the flue when taking pill #1, pill #2, or a placebo pill. For the null hypothesis, we would assume: $H_0: p_1 = p_2 = p_3$. That is, the proportion of people who get the flu is the same across all pills. difference is we do it w/ a 2way table!

Note: in section 11.1 we studied many proportions from independent samples and in the next section, 11.3, we will study many proportions from dependent samples.

Steps for Hypothesis Test for Homogeneity of Proportions

What to Find…

- Number of categories, k
- Expected Counts, E_i

$$E = (column\ total) \cdot \left(\frac{row\ total}{table\ total}\right)$$

Requirements:

- The sample data are randomly selected.
- The sample data are represented as frequency counts in a two-way table.
- For each cell in the contingency table, the expected frequency is at least 5.

Step 1: *Hypotheses*

H_0: the proportions are all equal

H_A: At least one proportion is different than the others

ALWAYS ___________-TAILED TEST!

Step 2: *Level of Significance*

Step 3: *Test Statistic*

$$\chi_0^2 = \sum \frac{(O-E)^2}{E} \quad \text{and} \quad df = (r-1)\cdot(c-1)$$

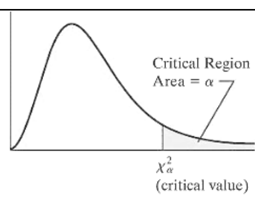
Step 4: *Find a Critical Value or P-Value to check using either the Critical Value or P-value method.*

Step 5: *Make a decision and draw a conclusion.*

CRITICAL REGION METHOD

* Table VIII

$df = (r-1)(c-1)$

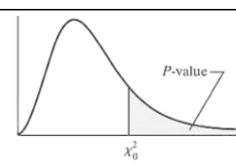


Reject H_0 if χ^{2*} lies in the critical region

Fail to Reject H_0 if χ^{2*} doesn't lie in the critical region

P-VALUE METHOD

$df = (r-1)(c-1)$



Reject H_0 if *P-value* $\le \alpha$

Fail to Reject H_0 if *P-value* $> \alpha$

GRAPHING CALCULATOR (TI-83 OR 84)

Instructions: (a) 2nd $\Rightarrow x^{-1}$ (matrix) $\Rightarrow$ EDIT $\Rightarrow$ 1: $[A]$ $\Rightarrow$ ENTER

(b) Input # of rows $\times$ # of columns ($r \times c$)

(c) Enter the sample values

(d) STAT $\Rightarrow$ TESTS $\Rightarrow$ C: χ^2-Test

Ex 3: Zocor is a drug manufactured by Merck and Co. that is meant to reduce the level of LDL (bad) cholesterol and increase the level of HDL (good) cholesterol. In clinical trials of the drug, patients were randomly divided into three groups. Group 1 received Zocor; group 2 received a placebo; group 3 received cholestyramine, a cholesterol-lowering drug that is currently available. The table contains the number of patients in each group who did and did not experience abdominal pain as a side effect.

Is there evidence to indicate that the proportion of subjects in each group who experienced abdominal pain is different at the $\alpha = 0.01$ level of significance?

2 way table

	Zocor	*Placebo*	*Cholestyramine*	
Abdominal Pain	51	5	16	$\Sigma =$
No Abdominal Pain	1532	152	163	$\Sigma =$
	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$

Find the Expected Frequencies

	Zocor	*Placebo*	*Cholestyramine*
Abdominal Pain			
No Abdominal Pain			

Check requirements

Null and Alternative Hypothesis

Test Statistic

P-Value

Decision about Null Hypothesis

Conclusion