SCSI2143: PROBABILITY & STATISTICAL DATA ANALYSIS

CHAPTER 6

Chi-Square Test & Contingency Analysis

(Chi-Square Test for *k* Proportions,
Chi-Square Test of Independence Contingency Table)

www.utm.my innovative • entrepreneurial • global 1

Chi-Square Test & One Way Contingency Table

- Categories with Equal Frequencies/Probabilities
- Categories with Unequal Frequencies/Probabilities

www.utm.my innovative • entrepreneurial • global 2

Multinomial Experiment

An experiment that meets the following conditions:

1. The number of trials is fixed.
2. The trials are independent.
3. All outcomes of each trial must be classified into exactly one of several different categories.
4. The probabilities for the different categories remain constant for each trial.

Multinomial Experiment (cont.)

- *n* identical trials
- *k* outcomes to each trial
- Constant outcome probability, p_k
- Independent trials
- Random variable is count, o_k
- Example: Ask 100 People (*n*) which of 3 candidates (*k*) they will vote for.

Goodness-of-fit Test

Goodness-of-fit test is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

Goodness-of-fit Test (cont.)

Notation:

O represents the observed frequency of an outcome

E represents the expected frequency of an outcome

k represents the number of different categories or outcomes

n represents the total number of trials

Expected Frequencies

If all expected frequencies are **equal**:

$$E = \frac{n}{k}$$

the sum of all observed frequencies divided by the number of categories.

Expected Frequencies (cont.)

If all expected frequencies are **not all equal**:

$$E = n*p$$

each expected frequency is found by multiplying the sum of all observed frequencies (n) by the probability for the category (p).

Expected Frequencies (cont.)

Key Question :

Are the differences between the **observed values (O)** and the theoretically **expected values (E)** statistically significant?

Answer:

We need to measure the discrepancy between O and E; the test statistic will involve their difference: O - E

Chi-Square Test

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Test statistic value-> calculated.

Critical Values (Chi-square value from table):

1. Found in table χ^2 using k-1 degrees of freedom

   where k = number of categories.

2. Goodness-of-fit hypothesis tests are always right-tailed.

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Test Hypothesis

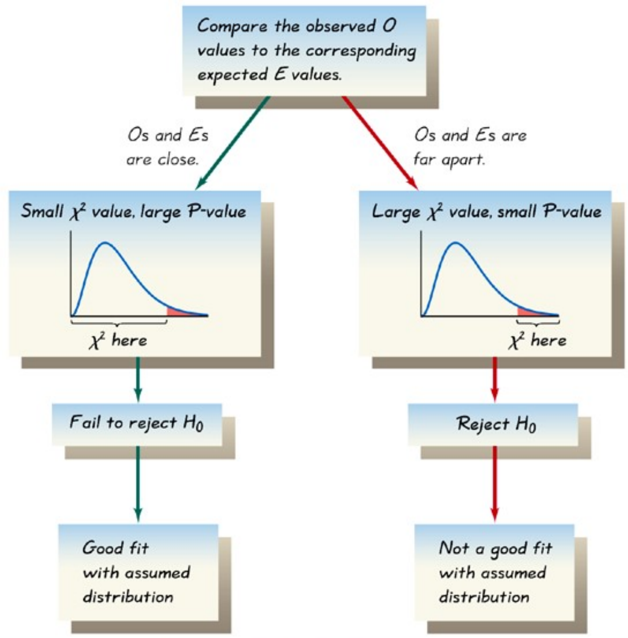
H_0: No difference between observed and expected probabilities.

H_1: At least one of the probabilities is different from the others.

UTM UNIVERSITI TEKNOLOGI MALAYSIA

- A **close agreement** between observed and expected values will lead to a small value of χ^2 and a large *p*-value.
- A **large disagreement** between observed and expected values will lead to a large value of χ^2 and a small *p*-value.
- A significantly large value of χ^2 will cause a rejection of the null hypothesis of no difference between the observed and the expected.

Relationships Among Components in Goodness-of-Fit Hypothesis Test



Chi-Square (χ^2) Test for *k* Proportions

- Tests Equality (=) of Proportions Only
  - Example: $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$
- One variable with several levels.
- Assumptions:
  - Multinomial Experiment
  - Large Sample Size
    - **All expected counts $\geq$ 5**
- Uses One-Way Contingency Table

UTM UNIVERSITI TEKNOLOGI MALAYSIA

One-Way Contingency Table

- Shows number of observations in *k* Independent Groups (Outcomes or Variable Levels)

Outcomes (k = 3)

Candidate			
Tom	Bill	Mary	Total
35	20	45	100

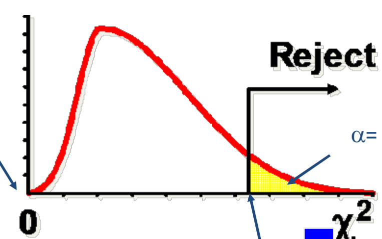
Number of responses

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Finding Critical Value

Example: What is the critical χ^2 value if k = 3, and α=0.05?

If $o_i = e_i$, $\chi^2 = 0$. Do not reject H_0



χ^2 Table (Portion)

	Upper Tail Area				
DF	.995	...	.95	...	.05
1	...	...	0.004	...	3.841
2	0.010	...	0.103	...	5.991

Df = k - 1 = 2

Categories With Equal Frequencies/Probabilities

Categories with Equal Frequencies/Probabilities

Statement of test hypothesis:

H_0: $p_1 = p_2 = p_3 = \ldots = p_k$

H_1: at least one of the probabilities is different from the others.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Example 1

A study was conducted on 147 cases of industrial accidents that required medical attention. Test the claim that the accidents occur with equal proportions on the 5 workdays.

Day	Mon	Tues	Wed	Thurs	Fri
Observed accidents	31	42	18	25	31

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Example 1 - Solution

Claim: Accidents occur with the same proportion. Therefore, $p_1 = p_2 = p_3 = p_4 = p_5$

i. State the test hypothesis:

H_0: $p_1 = p_2 = p_3 = p_4 = p_5$

H_1: At least 1 of the 5 proportions is different from others.

Example 1 – Solution (cont.)

ii. Calculate the expected frequency:

$$E = n/k = 147/5 = 29.4$$

Observed and Expected Frequencies

	Day	Mon	Tues	Wed	Thurs	Fri
O:	Observed accidents	31	42	18	25	31
E:	Expected accidents	29.4	29.4	29.4	29.4	29.4

Example 1 – Solution (cont.)

iii. Calculate the different between O and E:

$$(O - E)^2/E$$

Observed and Expected Frequencies of Industrial Accidents

Day	Mon	Tues	Wed	Thurs	Fri
Observed accidents	31	42	18	25	31
Expected accidents	29.4	29.4	29.4	29.4	29.4
$(O - E)^2/E$	0.0871	5.4000	4.4204	0.6585	0.0871 (rounded)

$$\frac{(O - E)^2}{E} = \frac{(31 - 29.4)^2}{29.4} = 0.0871$$

Example 1 – Solution (cont.)

Observed and Expected Frequencies of Industrial Accidents

Day	Mon	Tues	Wed	Thurs	Fri
Observed accidents	31	42	18	25	31
Expected accidents	29.4	29.4	29.4	29.4	29.4
$(O-E)^2/E$	0.0871	5.4000	4.4204	0.6585	0.0871 (rounded)

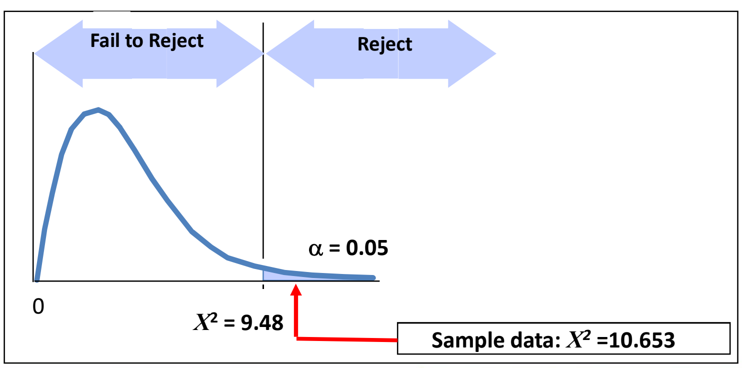
iv. Calculate the test statistic: (calculated chi-square value)

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 0.0871 + 5.4000 + 4.4204 + 0.6585 + 0.0871 = 10.6531$$

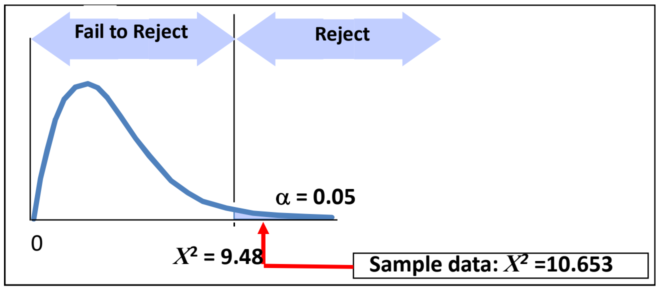
Example 1 – Solution (cont.)

v. Find the critical value: (chi-square value from table)

- Refer to table χ^2 with k-1 = 5 -1 = 4; and α = 0.05.
- It will show that $\chi^2_{4,0.05} = 9.48$



Example 1 – Solution (cont.)



vi. Based on the result, state the conclusion:

Test statistic falls within the critical region, therefore we reject hypothesis null. That is, we reject claim that the accidents occur with equal proportions (frequency) on the 5 workdays.

Example 2

As personnel director, you want to test the perception of fairness of three methods of performance evaluation.

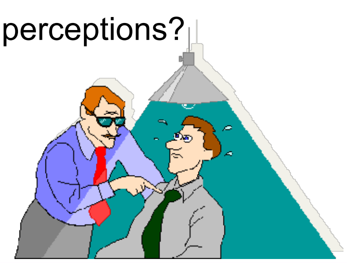
Of **180** employees,

63 rated **Method 1** as fair.

45 rated **Method 2** as fair.

72 rated **Method 3** as fair.

At the 0.**05** level, is there a **difference** in perceptions?



Example 2 – Solution (cont.)

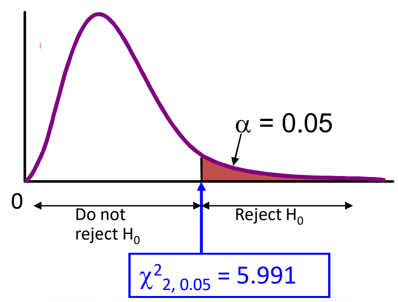
i. Test hypothesis:

H_0: $p_1 = p_2 = p_3 = 1/3$

H_1: At least 1 is different

ii. Find the critical value:

$\alpha = 0.05; k = 3-1 = 2$



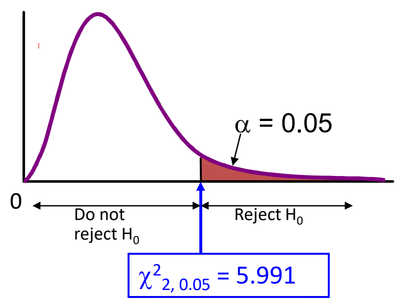
Example 2 – Solution (cont.)

iii. Calculate the expected counts and,
iv. Find the test statistics value:

Cell, i	Observed Count, o_i	Expected Count, e_i	$[o_i - e_i)]^2/ e_i$
1	63	$(1/3)\times180=60$	0.15
2	45	$(1/3)\times180=60$	3.75
3	72	$(1/3)\times180=60$	2.40
Total	180	180	$\chi^2 = 6.30$

Example 2 – Solution (cont.)

v. State the decision:



Test Statistic: $\chi^2 = 6.3$

Critical value: χ^2 (k=2, α=0.05)

= 5.991

Conclusion:

Reject H_0 at α = .05

There is evidence of a difference in proportions.

Example 3

- Are technical support calls equal across all days of the week?

  - Sample data:

	Sum of calls for each day:
Monday	290
Tuesday	250
Wednesday	238
Thursday	257
Friday	265
Saturday	230
Sunday	192
	Σ = 1722

Example 3 – Solution

- If calls **are** equal across all days of the week, the 1722 calls would be expected to be equally divided across the 7 days:

$$\frac{1722}{7} = 246 \quad \text{expected calls per day}$$

i. Test hypothesis:

H_0: $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 1/7$

H_1: At least 1 is different

Example 3 – Solution (cont.)

ii. Calculate the expected counts and,
iii. Find the test statistics value:

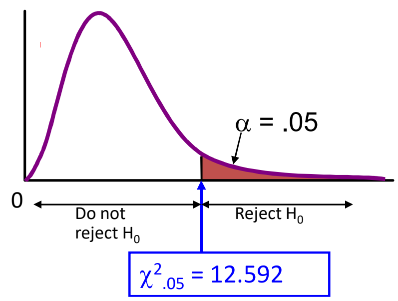
	Observed o_i	Expected e_i	$[o_i - e_i)]^2/ e_i$
Monday	290	246	7.8699
Tuesday	250	246	0.0650
Wednesday	238	246	0.2602
Thursday	257	246	0.4919
Friday	265	246	1.4675
Saturday	230	246	1.0407
Sunday	192	246	11.8537
TOTAL	1722	1722	χ^2 = 23.0489

Example 3 – Solution (cont.)

iv. Find the critical value:

- k – 1 = 6 (7 days of the week) so use 6 degrees of freedom

$\chi^2_{.05,6} = 12.592$



v. State the decision:

Conclusion:

$\chi^2 = 23.0489 > \chi^2_\alpha = 12.592$ so **reject H_0** and conclude that the distribution is not uniform.

Categories With Unequal Frequencies/Probabilities

Categories with Unequal Frequencies/Probabilities

H_0: p_1, p_2, p_3, . . . , p_k are as claimed.

H_1: at least one of the above proportions is different from the claimed value.

Example 4

Mars, Inc. claims its M&M candies are distributed with the color percentages of 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% blue. At the 0.05 significance level, test the claim that the color distribution is as claimed by Mars, Inc. The observed frequency as shown below:

Frequencies of M&Ms candies

	Brown	Yellow	Red	Orange	Green	Blue
Observed frequency	33	26	21	8	7	5

Example 4 - Solution

Claim: $p_{brown} = 0.30$, $p_{yellow} = 0.20$, $p_{red} = 0.20$,
$p_{orange} = 0.10$, $p_{green} = 0.10$, $p_{blue} = 0.10$

i. Statement of test hypothesis:

H_0 : $p_{brown} = 0.30$, $p_{yellow} = 0.20$, $p_{red} = 0.20$,
$p_{orange} = 0.10$, $p_{green} = 0.10$, $p_{blue} = 0.10$.

H_1: At least one of the proportions is different from the claimed value.

ii. Calculate the expected frequency:

Frequencies of M&Ms candies

	Brown	Yellow	Red	Orange	Green	Blue
Observed frequency	33	26	21	8	7	5

$n = 100$

Expected frequency:

Brown $E = np = (100)(0.30) = 30$

Yellow $E = np = (100)(0.20) = 20$

Red $E = np = (100)(0.20) = 20$

Orange $E = np = (100)(0.10) = 10$

Green $E = np = (100)(0.10) = 10$

Blue $E = np = (100)(0.10) = 10$

UTM UNIVERSITI TEKNOLOGI MALAYSIA

iii. Calculate the test statistic @chi-square value:

	Brown	Yellow	Red	Orange	Green	Blue
Observed frequency	33	26	21	8	7	5
Expected frequency	30	20	20	10	10	10
$(O - E)^2/E$	0.3	1.8	0.05	0.4	0.9	2.5

Test statistics value:

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 5.95$$

www.utm.my innovative • entrepreneurial • global 39

UTM UNIVERSITI TEKNOLOGI MALAYSIA

iv. Find the critical value from chi-square table:

Critical Value χ^2 =11.071
(with k-1 = 5 and α = 0.05)

v. State the decision:

Test statistic value (χ^2 =5.95) < critical value ($\chi^2_{k=5,\ \alpha=0.05}$ =11.071), that is it does not fall within critical region. Thus, we do not reject H_0.

There is not sufficient evidence to warrant rejection of the claim that the colors are distributed with the given percentages.

www.utm.my innovative • entrepreneurial • global 40

Exercise

It was claimed that population at ABC country in 2008 consisted of 50.7% English, 6.6% French, 30.6% Irish, 10.8% Asians, and 1.3% other ethnic groups. Suppose that a random sample of 1000 student graduating from ABC colleges and universities in 2008 resulted in the accompanying data on ethnic group (see table below).

Ethnic Group	Number in Sample
English	679
French	51
Irish	77
Asian	190
Other	3

Do the data provide evidence that the proportion of students graduating from colleges and universities in ABC for these ethnic group categories differs from the respective proportions in the population for ABC? Test the appropriate hypotheses using α=0.01.

Chi-Square Test & Two Way Contingency Table

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Chi-Square (χ^2) Test of Independence

- To shows if a relationship exists between 2 qualitative variables, when
  - One sample is drawn.
  - Does **not** show causality.
- Assumptions:
  - Multinomial experiment.
  - All expected counts ≥ 5
- Uses two-way contingency table

UTM UNIVERSITI TEKNOLOGI MALAYSIA

- Shows # observations from 1 sample jointly in 2 qualitative variables:

Level of variable: 2

House Style	House Location: Urban	House Location: Rural	Total
Split-Level	63	49	112
Ranch	15	33	48
Total	78	82	160

Level of variable: 1

Test hypotheses & Test Statistic

- Test hypothesis:
  H_0: Variables are independent.
  H_1: Variables are related (dependent).

- Test Statistic:

Observed count

Expected count

$$\chi^2 = \sum_{\text{all cells}} \frac{[o_{ij} - e_{ij}]^2}{e_{ij}}$$

Rows   Columns

- Degrees of Freedom: $(r - 1)(c - 1)$

Calculation of Expected Counts

- Statistical independence means joint probability equals product of marginal probabilities.
- Compute marginal probabilities & multiply for joint probability.
- Expected count is sample size times joint probability.

Example 5

House Style	Location Urban Obs.	Location Rural Obs.	Total
Split-Level	63	49	112
Ranch	15	33	48
Total	78	82	160

Example 5 (cont.)

Marginal probability = $\frac{112}{160}$

House Style	Location Urban Obs.	Location Rural Obs.	Total
Split-Level	63	49	112
Ranch	15	33	48
Total	78	82	160

Example 5 (cont.)

Marginal probability = $\frac{112}{160}$

House Style	Location: Urban Obs.	Location: Rural Obs.	Total
Split-Level	63	49	112
Ranch	15	33	48
Total	78	82	160

Marginal probability = $\frac{78}{160}$

Example 5 (cont.)

Joint probability = $\frac{112}{160} \times \frac{78}{160}$

Marginal probability = $\frac{112}{160}$

House Style	Location: Urban Obs.	Location: Rural Obs.	Total
Split-Level	63	49	112
Ranch	15	33	48
Total	78	82	160

Marginal probability = $\frac{78}{160}$

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Example 5 (cont.)

Expected Count calculation formula:

$$e_{ij} = \frac{(i^{th}\text{ Row total})(j^{th}\text{ Column total})}{\text{Total sample size}}$$

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Example 5 (cont.)

Joint probability = $\frac{112}{160} \times \frac{78}{160}$

Marginal probability = $\frac{112}{160}$

	Location		
House Style	Urban Obs.	Rural Obs.	Total
Split-Level	63	49	112
Ranch	15	33	48
Total	78	82	160

Marginal probability = $\frac{78}{160}$

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Example 5 (cont.)

	House Location				
	Urban		Rural		
House Style	Obs.	Exp.	Obs.	Exp.	Total
Split-Level	63	54.6	49	57.4	112
Ranch	15	23.4	33	24.6	48
Total	78	78	82	82	160

$\frac{112 \cdot 78}{160}$ $\frac{112 \cdot 82}{160}$ $\frac{48 \cdot 78}{160}$ $\frac{48 \cdot 82}{160}$

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Example 6

You're a marketing research analyst. You ask a random sample of **286** consumers if they purchase Diet Pepsi or Diet Coke. At the **0.05** level, is there evidence of a **relationship**?

	Diet Pepsi		
Diet Coke	No	Yes	Total
No	84	32	116
Yes	48	122	170
Total	132	154	286

Example 6 - Solution

i. State the test hypothesis:

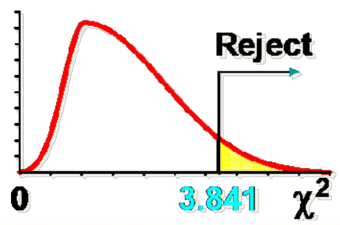
H_0: No relationship between variables.

H_1: Variables has relationship.

ii. Find the critical value (refer to chi-square table):

$\alpha = 0.05$

df = (2 - 1)(2 - 1) = 1



Example 6 – Solution (cont.)

iii. Calculate the expected counts:

	Diet Pepsi				
	No		Yes		
Diet Coke	Obs.	Exp.	Obs.	Exp.	Total
No	84	$\boxed{53.5}$	32	$\boxed{62.5}$	116
Yes	48	$\boxed{78.5}$	122	$\boxed{91.5}$	170
Total	132	132	154	154	286

$\frac{116 \cdot 132}{286}$ $\frac{116.154}{286}$ $\frac{170 \cdot 132}{286}$ $\frac{170 \cdot 154}{286}$

✓ $e_{ij} \geq 5$ in all cells

Example 6 – Solution (cont.)

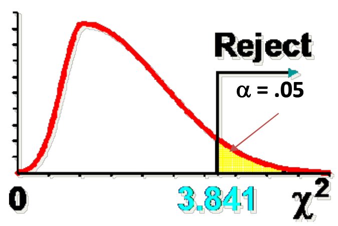
iv. Calculate the test statistic value:

Cell, ij	Observed Count, o_{ij}	Expected Count, e_{ij}	$[o_{ij} - e_{ij})]^2 / e_{ij}$
1,1	84	(116)(132)/286 =53.5	17.39
1,2	32	(116)(154)/286 =62.5	14.88
2,1	48	(170)(132)/286 =78.5	11.85
2,2	122	(170)(154)/286 =91.5	10.17
		$\chi^2=$	54.29

www.utm.my innovative • entrepreneurial • global 57

Example 6 – Solution (cont.)

v. State the decision:



Test Statistic: $\chi^2 = 54.29$

Critical value: $\chi^2_{k=1,\alpha=0.05} = 3.841$

Decision:

Since, test statistic value > critical value, thus reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence of a relationship between the variables.

www.utm.my innovative • entrepreneurial • global 58

Example 7

- Left-Handed vs. Gender
  - Dominant Hand: Left vs. Right
  - Gender: Male vs. Female

H_0: Hand preference is independent of gender
H_1: Hand preference is not independent of gender

Example 7 – Solution

- Sample results organized in a contingency table:

sample size = n = 300:

120 Females, 12 were left handed

180 Males, 24 were left handed

Gender	Hand Preference: Left	Hand Preference: Right	
Female	12	108	120
Male	24	156	180
	36	264	300

Example 7 – Solution (cont.)

- Observed frequencies vs. expected frequencies:

Gender	Hand Preference: Left	Hand Preference: Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

Example 7 – Solution (cont.)

Cell, ij	Observed Count, o_{ij}	Expected Count, e_{ij}	$[o_{ij} - e_{ij})]^2 / e_{ij}$
1,1	12	(120)(36)/300 =14.4	0.4000
1,2	108	(120)(264)/300 =105.6	0.0545
2,1	24	(180)(36)/300 =21.6	0.2667
2,2	156	(180)(264)/300 =158.4	0.0364
		χ^2=	0.7576

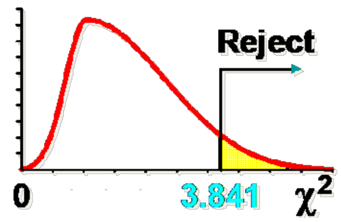
Example 7 – Solution (cont.)

Test Statistic: $\chi^2 = 0.7576$

Critical value: $\chi^2_{k=1,\alpha = 0.05} = 3.841$

Decision:

Since, test statistic value < critical value, thus do not reject H_0 at $\alpha = 0.05$



Conclusion:

There is evidence that gender and hand preference are independent.

Exercise

Jail inmates can be classified into one of the following four categories according to the type of crime committed: violent crime, crime against property, drug offenses, and public-order offenses. Suppose that random samples of 500 male inmates and 500 female inmates are selected, and each inmate is classified according to type of offense.

Type of Crime	Gender	
	Male	Female
Violent	117	66
Property	150	160
Drug	109	168
Public-order	124	106

We would like to know whether male and female inmates differ with respect to type of offense. Test the relevant hypotheses using a significance level of 0.05.