

Supply Chain Management 651 Fall 2017 (SCM 651)
Whitman School of Management
Lecture Notes for Week 3

Coverage Session 3:

1. Review: Buy response curve, Excel: sort, filter

We will use Excel Data for Session 3

2. Chi-Square Analysis with cross-tabulations (Chapter 1 of Course Reader)

We will use Carrier Dome Data as example.

3. Review group assignment 1

Coverage of Session 4: Regression analysis (Chapter 2 of Course Reader)

Coverage of Session 5: Conjoint analysis, Logit (Chapters 3 and 5 of Course Reader)

Chi-Square Test with Cross-Tabulation (Chapter 1 of Course Reader)

Pivot-table: For two variables V_1 and V_2 , each with a finite number of categories, a pivot table is a table where each cell corresponds to the same combination of values of V_1 and V_2 . Once the table is created, a statistic about a third variable computed that particular combination of values of V_1 and V_2 is entered in each cell.

Example: For a city, V_1 is the neighborhood where a house is located (East, West or North), and V_2 is the type of house (not brick, or brick). For each of six combinations of V_1 and V_2 (for example, East and brick), the entry may be average assessed value of the house.

Cross-tabulation: This is a special case of a pivot table where the entry in each cell is the number of times that combination of V_1 and V_2 occurs in the sample.

Chi-square test with Cross-tabulation: For a cross-tabulation of two variables V_1 and V_2 obtained from a sample, chi-square analysis is used to test the null hypothesis that V_1 and V_2 are **not related** to each other.

Learning Objectives:

- Meaning of “no relationship.”
- The chi-square test
 - Compute expected frequencies.
 - Compute chi-square (χ^2).
 - Compute degrees of freedom.
 - Do the test.
- Check if test is valid and combine rows and/or columns as necessary to have a valid test.
- Important application: Test if population proportion is same in two or more sub-populations.

Example 1.2 We selected a simple random sample of 150 students from a college campus, and recorded (i) the gender of the student, and (ii) whether the student has attended a basketball game played by the college team during the past year. The results are expressed as the following 2×2 cross tabulation:

	Didn't Attend Game	Attended Game
Male	30	60
Female	42	18

H_0 : There is no relationship between gender and attendance.

Intuition of H_0 :

- There are two sub-populations: men and women.
- A member of either sub-population belongs to one of two categories: did not attend, and attended.
- If H_0 is true, then each sub-population (men or women) should have the same percentage break-down between the two categories of attendance.

Formally: Define:

π_{11} = Proportion of men who did not attend	π_{12} = Proportion of men who attended
π_{21} = Proportion of women who did not attend	π_{22} = Proportion of women who attended

H_0 means: $\pi_{11} = \pi_{21}$, $\pi_{12} = \pi_{22}$.

	Didn't Attend Game	Attended Game
Male	30	60
Female	42	18

H_0 : There is no relationship between gender and attendance.

Expected Frequencies: In the whole sample of 150 students:

$$\text{Proportion that did not attend} = \frac{\text{Total of Column 1}}{n} = \frac{72}{150}$$

$$\text{Proportion that attended} = \frac{\text{Total of Column 2}}{n} = \frac{78}{150}$$

$$\text{Number of men in sample} = \text{Total of Row 1} = 90$$

Expected number of men that did not attend (E_{11})

$$= 90 \times \frac{72}{150} = \frac{90 \times 72}{150} = 43.2$$

Expected number of men that attended (E_{12})

$$= 90 \times \frac{78}{150} = \frac{90 \times 78}{150} = 46.8$$

$$\text{Number of women in sample} = \text{Total of Row 2} = 60$$

Expected number of women that did not attend (E_{21})

$$= 60 \times \frac{72}{150} = \frac{60 \times 72}{150} = 28.8$$

Expected number of women that did attended (E_{22})

$$= 60 \times \frac{78}{150} = \frac{60 \times 78}{150} = 31.2$$

$$\text{More Generally: } E_{ij} = \frac{\text{Total of Row } i \times \text{Total of Column } j}{\text{Sample Size}(n)}$$

$$E_{ij} = \frac{\text{Total of Row } i \times \text{Total of Column } j}{\text{Sample Size}(n)}$$

Chi-Square:

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

(Compute $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ in each cell, and then sum over all cells.)

R = number of rows

C = number of columns

Degrees of freedom = $(R - 1) \times (C - 1)$

Decision Rule: At a confidence level $(1 - \alpha)$, reject H_0 if $\chi^2 > \chi^2_{\alpha}$ at degree of freedom $(R - 1) \times (C - 1)$

Alternative Decision Rule

- At a $(1 - \alpha)$ level of confidence, reject H_0 if P value is less than α .
- P value = probability that chi-square at degree of freedom $(R - 1) \times (C - 1)$ exceeds the computed chi-square.
- Suppose cell A1 in your Excel worksheet has the computed chi-square, and cell A2 in your Excel worksheet has the degree of freedom. Then, in any cell, type =CHISQ.DIST.RT(A1,A2) and hit enter. Excel will return the P value.

Degrees of freedom = $(R - 1) \times (C - 1)$

Decision Rule: At a confidence level $(1 - \alpha)$, reject H_0 if $\chi^2 > \chi^2_\alpha$ at degree of freedom $(R - 1) \times (C - 1)$

Return to Example 1.2 We selected a simple random sample of 150 students from a college campus, and recorded (i) the gender of the student, and (ii) whether the student has attended a basketball game played by the college team during the past year. The results are expressed as the following 2×2 cross tabulation:

	Didn't Attend Game	Attended Game
Male	30	60
Female	42	18

At a 95% level of confidence, test the null hypothesis that there is no relationship between gender and attendance (against the alternate hypothesis that there is some kind of relationship between the two).

Here:

$$O_{11} = \quad , \quad E_{11} = \frac{\quad \times \quad}{\quad} = \quad , \quad O_{12} = \quad , \quad E_{12} = \frac{\quad \times \quad}{\quad} =$$

$$O_{21} = \quad , \quad E_{21} = \frac{\quad \times \quad}{\quad} = \quad , \quad O_{22} = \quad , \quad E_{22} = \frac{\quad \times \quad}{\quad} =$$

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$

Degrees of freedom = $(2 - 1) \times (2 - 1) =$

Decision Rule: At a 95% level of confidence, reject H_0 if $\chi^2 >$

Conclusion:

Example 1.3 A random sample of 100 students was drawn the students of Syracuse University. For each student, the following information were recorded:

- Gender: female or male
- Interest in shopping for clothes

Results:

Gender	Interest		
	Low	Medium	High
Female	9	12	39
Male	17	10	13

At a 99% level of confidence, test the null hypothesis that there is no relationship between gender and interest in shopping for clothes.

(1) Write a new table by adding row totals and column totals:

Gender	Interest			Row Totals
	Low	Medium	High	
Female	9	12	39	
Male	17	10	13	
Column Totals				$n = 100$

Observed frequencies:

$$O_{11} = 9 \quad O_{12} = 12 \quad O_{13} = 39$$

$$O_{21} = 17 \quad O_{22} = 10 \quad O_{23} = 13$$

Expected frequencies:

$$E_{11} = \frac{\quad \times \quad}{\quad} = \quad \quad E_{12} = \frac{\quad \times \quad}{\quad} = \quad \quad E_{13} = \frac{\quad \times \quad}{\quad} = \quad$$

$$E_{21} = \frac{\quad \times \quad}{\quad} = \quad \quad E_{22} = \frac{\quad \times \quad}{\quad} = \quad \quad E_{23} = \frac{\quad \times \quad}{\quad} = \quad$$

$$\chi^2 =$$

Gender	Interest		
	Low	Medium	High
Female	9	12	39
Male	17	10	13

$$\chi^2 =$$

Number of rows $R =$, number of columns $C =$, degrees of freedom =
 $(R - 1) \times (C - 1) =$

Decision Rule: At a 99% level of confidence, reject H_0 if

$$\chi^2 > \quad = \chi^2_{.01} \text{ at degrees of freedom } = \quad .$$

Conclusion:

Validity of Chi-Square Test

Need:

- $E_{ij} > 1$ in **all** cells.
- $E_{ij} \geq 5$ is 80% or more of the cells.

Note:

- We may combine any rows or columns. If we have scale variables (e.g., 1-7 scales), combine adjacent rows or columns for ease of interpretation.
- If we combine any two rows or any two columns, the observed numbers add up, the expected numbers also add up.
- If you need to modify a table, any valid modification is acceptable. In Minitab you do that by recoding variables.
- The final modified table must have at least two rows and at least two columns.
- After you modify the table, the degree of freedom comes from the number of rows and columns of the modified table.

Example 1.4 We collected a simple random sample of 60 students from a college campus and asked them to rate how much they like to watch professional sports on TV on a 1-7 scale (strongly dislike to strongly like). We also noted the gender of each respondent. Based on the results, we have constructed the following cross-tabulation:

Gender	Like to watch professional sports on TV						
	1	2	3	4	5	6	7
Male	2	0	4	12	6	8	8
Female	4	3	2	6	3	1	1

At a 99% level of confidence, test the null hypothesis that gender is not related to how much one likes to watch professional sports on TV.

Process: First augment the cross-tabulation by row totals and column totals:

Gender	Like to watch sports on TV							Row Totals
	1	2	3	4	5	6	7	
Male	2	0	4	12	6	8	8	40
Female	4	3	2	6	3	1	1	20
Column Totals	6	3	6	18	9	9	9	

Then compute the expected frequencies in the original table.

$E_{11} = \frac{40 \times 6}{60} =$	$E_{12} = \frac{40 \times 3}{60} =$	$E_{13} = \frac{40 \times 6}{60} =$	$E_{14} = \frac{40 \times 18}{60}$ =
$E_{15} = \frac{40 \times 9}{60} =$	$E_{16} = \frac{40 \times 9}{60} =$	$E_{17} = \frac{40 \times 9}{60} =$	
$E_{21} = \frac{20 \times 6}{60} =$	$E_{22} = \frac{20 \times 3}{60} =$	$E_{23} = \frac{20 \times 6}{60} =$	$E_{24} = \frac{20 \times 18}{60}$ =
$E_{25} = \frac{20 \times 9}{60} =$	$E_{26} = \frac{20 \times 9}{60} =$	$E_{27} = \frac{20 \times 9}{60} =$	

Is it valid to use chi-square test with original table?

$E_{11} = \frac{40 \times 6}{60} = 4$	$E_{12} = \frac{40 \times 3}{60} = 2$	$E_{13} = \frac{40 \times 6}{60} = 4$	$E_{14} = \frac{40 \times 18}{60} = 12$
$E_{15} = \frac{40 \times 9}{60} = 6$	$E_{16} = \frac{40 \times 9}{60} = 6$	$E_{17} = \frac{40 \times 9}{60} = 6$	
$E_{21} = \frac{20 \times 6}{60} = 2$	$E_{22} = \frac{20 \times 3}{60} = 1$	$E_{23} = \frac{20 \times 6}{60} = 2$	$E_{24} = \frac{20 \times 18}{60} = 6$
$E_{25} = \frac{20 \times 9}{60} = 3$	$E_{26} = \frac{20 \times 9}{60} = 3$	$E_{27} = \frac{20 \times 9}{60} = 3$	

Writing Compactly, the Expected Frequencies (E_{ij} 's) are:

	Like to watch professional sports on TV						
Gender	1	2	3	4	5	6	7
Male	4	2	4	12	6	6	6
Female	2	1	2	6	3	3	3

Items to check:

- Is $E_{ij} > 1$ in all cells?
- If yes, then is $E_{ij} \geq 5$ in 80% or more cells?

Note:

- If either condition fails, you have to combine columns to get a valid test.
- Since there are only two rows, you cannot combine rows here.
- If you had more than two rows, you could have combined rows. (Note that the final table must have at least two rows and at least two columns.)

Old Table of Observed Frequencies (O_{ij} 's):

	Like to watch professional sports on TV						
Gender	1	2	3	4	5	6	7
Male	2	0	4	12	6	8	8
Female	4	3	2	6	3	1	1

Old Table of Expected Frequencies (E_{ij} 's):

	Like to watch professional sports on TV						
Gender	1	2	3	4	5	6	7
Male	4	2	4	12	6	6	6
Female	2	1	2	6	3	3	3

New Table:

Observed and Expected Frequencies:

Degrees of freedom = (-1) \times (-1) =
 $\chi^2 =$

Decision Rule: At a 99% level of confidence, reject H_0 if $\chi^2 >$

Conclusion:

An Important Application: Using Chi-Square Analysis to test if proportions are equal in multiple sub-populations

Example 1.5 Suppose you have collected simple random samples from three sub-populations: Business majors, Engineering majors, and “other” majors. For each respondent, you recorded if (s)he reads the Wall Street Journal (WSJ) every week. **Results:**

- (1) Business sample: $n_1 = 100$, 50 read WSJ every week.
- (2) Engineering sample: $n_2 = 50$, 15 read WSJ every week.
- (3) “Other: sample: $n_3 = 150$, 25 read WSJ every week.

At a 99% level of confidence, test the null hypothesis that an equal proportion of business, engineering, and other students read WSJ every week.

Approach: Express as a cross-tabulation:

	Do not Read	Read
Business	50	50
Engineering	35	15
Other	125	25

Use chi-square test to test the null hypothesis that there is no relationship between major and reading WSJ every week.

Logic: Let:

- π_{11} = fraction of business majors who do not read WSJ every week
- π_{12} = fraction of business majors who read WSJ every week
- π_{21} = fraction of engineering majors who do not read WSJ every week
- π_{22} = fraction of engineering majors who read WSJ every week
- π_{31} = fraction of other majors who do not read WSJ every week
- π_{32} = fraction of other majors who read WSJ every week

Clearly:

- $\pi_{11} + \pi_{12} = 1$, that is, $\pi_{11} = 1 - \pi_{12}$
- $\pi_{21} + \pi_{22} = 1$, that is, $\pi_{21} = 1 - \pi_{22}$
- $\pi_{31} + \pi_{32} = 1$, that is, $\pi_{31} = 1 - \pi_{32}$

Therefore, if $\pi_{12} = \pi_{22} = \pi_{32}$, we also have $\pi_{11} = \pi_{21} = \pi_{31}$

Denoting proportions of business, engineering and other majors that read WSJ by π_1 , π_2 and π_3 , respectively, and comparing definitions, we have:

$$\pi_1 = \pi_{12}, \pi_2 = \pi_{22}, \text{ and } \pi_3 = \pi_{32}$$

Therefore, the following two null hypotheses are equivalent:

(1) $\pi_1 = \pi_2 = \pi_3$, that is, equal proportions of the three sub-populations read WSJ every week.

(2) $\pi_{11} = \pi_{21} = \pi_{31}$ **and** $\pi_{12} = \pi_{22} = \pi_{32}$, that is, there is no relationship between major and reading WSJ every week.

Example 1.5 (continued):

	Do not Read	Read	Row Totals
Business	50	50	
Engineering	35	15	
Other	125	25	
Column Totals			Sample Size = 300

$$E_{11} = \frac{\quad \times \quad}{\quad} = \quad \quad E_{12} = \frac{\quad \times \quad}{\quad} =$$

$$E_{21} = \frac{\quad \times \quad}{\quad} = \quad \quad E_{22} = \frac{\quad \times \quad}{\quad} =$$

$$E_{31} = \frac{\quad \times \quad}{\quad} = \quad \quad E_{32} = \frac{\quad \times \quad}{\quad} =$$

$$\begin{aligned} \chi^2 &= \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} \\ &+ \frac{(O_{31} - E_{31})^2}{E_{31}} + \frac{(O_{32} - E_{32})^2}{E_{32}} \\ &= \frac{(\quad - \quad)^2}{\quad} + \frac{(\quad - \quad)^2}{\quad} + \frac{(\quad - \quad)^2}{\quad} + \frac{(\quad - \quad)^2}{\quad} \\ &+ \frac{(\quad - \quad)^2}{\quad} + \frac{(\quad - \quad)^2}{\quad} = \end{aligned}$$

Decision Rule: At a 99% level of confidence, reject H_0 if $\chi^2 > \chi_{.01}^2$ at $df = (3 - 1) \times (2 - 1) = 2$

Conclusion:

More Generally:

- Suppose you are testing if an equal proportion of k sub-populations have a property of interest (e.g., read Wall Street Journal every week), that is,

$$\pi_1 = \pi_2 = \dots = \pi_k$$

- This is equivalent to a chi-square test with a $k \times 2$ cross-tabulation where each row comes from one sub-population, and the two columns are “do not have property,” and “have property.”
- Express the data as a $k \times 2$ cross tabulation. For any sub-population:

Number who do not have property

= Size of the sample from the sub-population – Number from sub-population who have property

- Assuming test is valid, reject H_0 if χ^2 exceeds χ_α^2 at degrees of freedom $(k - 1) \times (2 - 1) = k - 1$.

Chi-Square Analysis with Minitab

We use Carrier Dome Data posted on Blackboard.

X_1 : Gender (1 if male, 0 if female)

X_{10b} : 1 if the student attended a football game at the Carrier Dome. (X_{10a} is same for basketball.)

$X_{8a} - X_{8l}$: Interest in activities in leisure time (1: not interested at all, 7: very interested).

a: exercise, b: participate in sports, c: shop for clothes, d: go to bars, e: go to malls, f: watch movies, g: do volunteer work, h: study/read, i: listen to music, j: spend time with friends, k: watch sports on TV, l: watch sports at carrier Dome

Example 1. Suppose you wish to test if there is a relation between gender (X_1), and if the student attended a football game in the Carrier Dome in the “last one year” (X_{10b}). Proceed as follows:

- (1) Open the worksheet in Minitab.
- (2) Click “Stat” in the menu line, drag cursor to “Tables,” and click on “Cross Tabulation and Chi-Square.” A dialog box opens.
- (3) Click in the top line on the right (“Rows”). Click on X_1 in the list in the left box and then click “Select.”
- (4) Mark the second line from the top on the right (“Columns”). Click on X_{10b} in the left box and then click “Select.”
- (5) Under “Display” at the right of the screen, click “Counts” and “Row percents.”
- (6) In the right of the dialog box, click on “Chi-square.” In the box that opens, mark “Chi-square test,” and click OK.
- (7) Back in main dialog box, click OK.

Minitab Output:

Rows X_1	Columns X_{10b}		
	0	1	All
0	38 50.67	37 49.33	75 100.00
1	14 19.18	59 80.82	73 100.00
All	52 35.14	96 64.86	148 100.00

Pearson Chi-Square = 16.095, DF = 1, P-Value = 0.000

The output includes:

- The cross-tabulation itself is:

X_1	X_{10b}	
	0	1
0	38	37
1	14	59

- The column at the right of the output table gives the row totals and sample size. The row at the bottom of the output table gives column totals and sample size.
- The percentage of each row ($X_1 = 0$ and $X_1 = 1$) that has $X_{10b} = 0$ and $X_{10b} = 1$. Thus, 49.33% of women and 80.82% men attended a football game at Carrier Dome.
- The computed Chi-Square.
- The degree of freedom (DF) = $(2 - 1) \times (2 - 1) = 1$.
- The P value for the null hypothesis of no relationship. Here, $P = 0.000$ means the P value is less than 0.001. Thus, H_0 can be rejected at a 99.9% level of confidence.

Note: If you wish to get the expected frequencies (E_{ij} 's) in the cells, mark "Expected cell counts" in addition to "Chi-square analysis" in Step 6 above.

Example 2. Suppose you wish to test if there is a relation between gender (X_1) and the student's interest in going to bars (X_{8d}). Proceeding as in Example 1, you get the following output:

Rows: X1	Columns: X8d							
	1	2	3	4	5	6	7	All
0	0	2	4	8	14	19	28	75
	0.00	2.67	5.33	10.67	18.67	25.33	37.33	100.00
1	1	1	3	8	17	21	22	73
	1.37	1.37	4.11	10.96	23.29	28.77	30.14	100.00
All	1	3	7	16	31	40	50	148
	0.68	2.03	4.73	10.81	20.95	27.03	33.78	100.00

Pearson Chi-Square = 2.560, DF = 6

* WARNING * 2 cells with expected counts less than 1

* WARNING * Chi-Square approximation probably invalid

* NOTE * 6 cells with expected counts less than 5

- Here, chi-square analysis is not valid with original table.
- As there are only two rows, we must combine columns.
- The problem arises because columns 1, 2, and 3 have too few observations. You can also do this step formally by looking at expected frequencies in the cells (E_{ij} 's). To get expected frequencies, do the following:

After you click “Chi-Square” (Step 6 in Example 1), mark both “Chi-Square test” and “Expected cell counts” and then click OK.

Chi-Square Analysis with Recoded Data

Example of recoding

Suppose we wish to create a variable BAR which is 1 if $X_{8d} = 1, 2, 3$; 2 if $X_{8d} = 4$; 3 if $X_{8d} = 5, 6$, or 7. We can create the recoded variable as follows:

- Click on “Data,” then “Code,” then “to Numeric.”
- Mark the “Code values in the following columns” in the top dialog box and select X_{8d} .
- In the line “Method,” use drop down menu and click on “Code range of values.” You will get a table to fill in values. Fill up the table as follows.

Lower endpoint	Upper endpoint	Coded value
1	3	1
4	4	2
5	7	3

- Using the drop down menu, change “end points to include” to “Both end-points.”
- Click OK. The recoded variable will be placed at the right of the current worksheet.
- In the cell at the top of the column of the recoded variable, enter BAR .

You can now do a cross-tabulation of X_1 and BAR . The results are as follows:

Rows: X1	Columns: BAR2			
	1	2	3	All
0	6	8	61	75
	8.00	10.67	81.33	100.00
1	5	8	60	73
	6.85	10.96	82.19	100.00
All	11	16	121	146
	7.43	10.81	81.76	100.00

Pearson Chi-Square = 0.072, DF = 2, P-Value = 0.965

We cannot reject H_0 at any reasonable level of confidence.

Chi-Square Test with Excel

We illustrate how to perform chi-square test using Excel in a separate document. A sketch is provided here.

- Open the data as an Excel worksheet and use pivot table to construct the cross tabulation of the two variables. Copy the cross tabulation to a new worksheet.
- Note that the cross tabulation already has row totals, column totals, and sample size computed. Use these to compute expected frequencies in new cells in the same order as the expected frequencies. For example, if the observed frequencies are in cells B2:H5, the expected frequencies may be in cells B8:H11.
- In any blank cell, type `chitest(observed range, expected range)` and hit enter. For example, `chitest(B2:H5,B8:H11)` and hit enter. Excel returns the P value.

More Examples

The Excel data for session 3 will be used in this class and also later when we discuss regression and logit. The data have the following work-sheets. Five of the work-sheets are adapted from material made available by Professor Johannes Ledolter for use with his text on business analytics.

1. The first worksheet is called “Death Penalty Ledolter.” This dataset provides 302 records of murder trials where death penalty was considered. The data include the fields:

- $D = 1$ if the victim is white, $D = 0$ if the victim is not white.
- X = aggravation factor measured on 1 to 6 (very low to very high) scale.
- $DX = D * X$
- $Y = 1$ if death penalty was awarded, 0 if not.

2. The second worksheet is called “Flight Delays Ledolter.” The data include 2198 records of flights with origins at BWI, DCA or IAD, and destination EWR, JFK, or LGA. Data fields are:

scheduled time	carrier	departure time
destination (EWR, JFK, or LGA)	distance	date
flight number	origin (BWI, DCA or IAD)	weather (0 or 1)
day of week	day of month	tail number
delay	delaynew (0: on time, 1: delayed)	

3. The third worksheet is called “Salary Discrimination Ledolter.” This data set gives the gender, number of years of experience, and salary of a sample of workers. Three new variables are created:

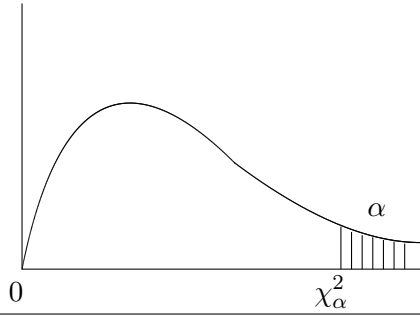
- $D = 1$ if the worker is male, and $D = 0$ if female.
- X is the number of years of experience.
- $DX = D * X$

4. The fourth worksheet, called “Male”, is the data from the third worksheet (salary, and X = years of experience) only for male workers.
5. The fifth worksheet, called “Female,” is the data from the third worksheet (salary, and X = years of experience) only for female workers.
6. The sixth worksheet, called “Titanic,” is the survival data from the Titanic disaster.
7. The seventh worksheet explains some terms in the Titanic worksheet.
8. The eighth worksheet is a sample of 3000 cases drawn from the Dominicks data base on the weekly sales of three brands of orange juice (HH, Minute Maid, Tropicana).

Note: I plan to do the following chi-square tests in the class:

- Using the Death Penalty Data Ledolter worksheet, test the null hypothesis of no relationship between
 - Victim’s ethnicity (D) and whether death penalty was awarded (Y).
 - Aggravation factor (X) and whether death penalty was awarded.
- Using the Flight Delay Ledolter worksheet, test the null hypothesis of no relationship between
 - Origin and whether the flight was delayed (delaynew)
 - Destination and whether the flight was delayed (delaynew)
- Using Titanic worksheet, test the null hypothesis of no relationship between
 - Passenger class (pclass) and whether the passenger survived (survived)
 - Gender and whether the passenger survived (survived).

Table 2. Chi-Square (χ^2) Distribution



df	α				
	.10	.05	.025	.01	.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.95
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.73	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
21	29.62	32.67	35.48	38.93	41.40
22	30.81	33.92	36.78	40.29	42.80
23	32.01	35.17	38.08	41.64	44.18
24	33.20	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.19	46.96	49.65
28	37.92	41.34	44.46	48.28	50.99
29	39.09	42.56	45.72	49.59	52.34
30	40.26	43.77	46.98	50.89	53.67

SCM 651 Fall 2017: Material Covered in Sessions 1 and 2

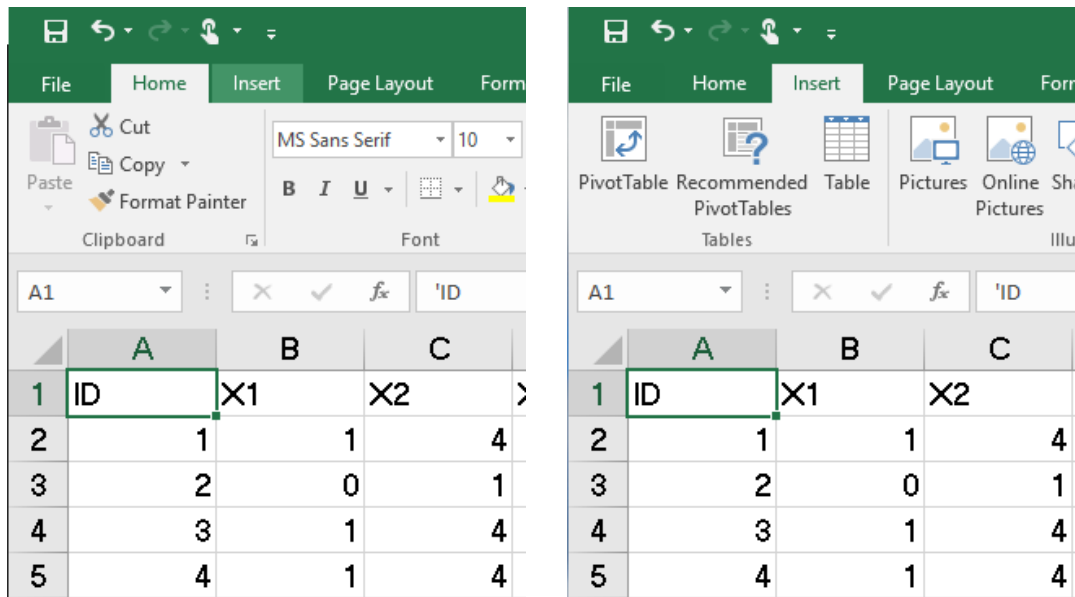
Topic	Relevance	Will I need it later in the course?
Charts and pivot charts	These are useful in presenting findings.	These will not be included in tests or assignments.
Pivot Table	Pivot tables are used to prepare summary statistics (such as mean, standard deviation) as well as cross tabulations. Also, using filters, you can focus on subsets of data. These can be useful by themselves, or with other methods (such as chi-square test with a cross tabulation).	You need pivot tables in assignments and the data analysis exam at the end of the semester.
Scatter Plots and trend-lines: <ul style="list-style-type: none"> • Linear plot (default) • Power curve • Exponential curve • Moving average 	<p>The linear plot can be used to estimate coefficients in two-variable (simple) regression.</p> <p>Power curve is applicable in two contexts: The experience curve, and constant elasticity demand models.</p> <p>Exponential curve is useful in modeling wealth under compound interest and capital growth. It can also be used to model decay (such as radioactive decay, sales decay if there is no advertising, forgetting).</p> <p>Moving average: This is used to remove the effects of seasonal variation to reveal trends in sales.</p>	<ul style="list-style-type: none"> • We will use linear plots throughout the course. For example, we need them in group assignment 1. • The idea of the power curve is important as we will use the constant elasticity demand model throughout the course. However, we will use packages such as R to estimate the regression models and not rely on scatter plots. • Exponential model, while important, will not be there in assignments or tests. • Moving average will be used in exploration to see if we need to include seasonal variation in regression models.
Data analysis tool pack	Often this is the only data analysis package available to you.	We need it in group assignment 1. After that, we will use packages such as R.
IRR, NPV, XNPV, CLV	<ul style="list-style-type: none"> • NPV is useful when comparing alternative investment plans that generate cash flows over time. The function assumes equal time intervals. 	NPV and IRR may be included in the data analysis exam. There will be no

	<ul style="list-style-type: none"> • IRR is the compound interest rate at which a project becomes viable. • XNPV allows you to compute net present value when cash flows happen at unequal time intervals. • CLV uses the same idea as NPV by combining interest rate and retention rate. It can tell you what the minimum retention rate should be to cover the cost of getting a new customer. 	question on CLV in tests or assignments.
Sorting and filtering	Organizing data	We use these throughout the course. Later on, we will use Access to do these tasks.
Buy response curve	Prepare demand function from survey data.	You need it in group assignment 1. There may be questions like this in the data analysis exam also.
What-if Analysis: Data Table	This is useful when you have a complex objective function, and you want to see how it varies with decision variables or model parameters. You can vary one or two variables/parameters at the same time.	You need it in group assignment 1.
What-if Analysis: Goal-seek	You can find the value of a variable you need to make an outcome take a given value. For example, you can change interest rate so that NPV is equal to zero (this gives you IRR).	You should know the method. However, there will be no assignment or test questions on goal-seek.
Conditional Formatting	This allows you to identify regions where an objective function is high or low.	While there is no assignment or test question on this, this is very useful when making presentations.
Solver	We have already used Solver to find IRR and profit maximizing price. We will discuss Solver more thoroughly later in the course.	There is an assignment on Solver (Assignment 3).

SCM 651 Fall 2017: Chi-square Test with Excel

Chi-Square Test: You can perform chi-square test using Excel pivot tables without using the data analysis add-in. I will show you how to do the test using the Carrier Dome data and the null hypothesis that there is no relationship between X1 (gender) and X8c (interested in shopping for clothes). Proceed as follows:

1. Open the Carrier Dome work-sheet. In the menu ribbon, click insert → Pivot Table
In the dialog box “Create Pivot Table,” click OK



	A	B	C
1	ID	X1	X2
2	1	1	4
3	2	0	1
4	3	1	4
5	4	1	4

Create PivotTable

Choose the data that you want to analyze

☒ Select a table or range

Table/Range: carrier_dome!\$A\$1:\$CC\$149

☐ Use an external data source

Choose Connection...

Connection name:

☐ Use this workbook's Data Model

Choose where you want the PivotTable report to be placed

☒ New Worksheet

☐ Existing Worksheet

Location:

Choose whether you want to analyze multiple tables

☐ Add this data to the Data Model

OK Cancel

- In the create table worksheet, drag X1 to rows, X8c to Columns, and ID to Σ Values. In the Σ values box, click Sum of ID \rightarrow Value Field Settings \rightarrow Count

This gives you the cross tabulation of X1 and X8c.

The image shows two parts of the Excel interface. On the left is the 'Value Field Settings' dialog box. It has a 'Source Name' of 'ID' and a 'Custom Name' of 'Count of ID'. Under 'Summarize Values By', the 'Show Values As' tab is selected. In the 'Summarize value field by' section, a list of calculation types is shown with 'Count' selected. On the right is a preview of the PivotTable layout. It shows 'X8c' in the COLUMNS area and 'Sum of ID' in the Σ VALUES area. The ROWS area contains 'X1'. There are buttons for 'Defer Layout Update' and 'UPDATE'.

Count of ID									
	A	B	C	D	E	F	G	H	I
1	Drop Report Filter Fields Here								
2									
3	Count of ID	X8c							
4	X1		1	2	3	4	5	6	7 Grand Total
5		0	3	4	4	14	17	18	15 75
6		1	7	11	13	19	16	6	1 73
7	Grand Total		10	15	17	33	33	24	16 148

- Computing Expected Frequencies: The pivot table with count of ID in Σ VALUES is the cross-tabulation of X1 and X8c and gives you the observed frequencies. You now need to compute expected frequencies for the cross-tabulation. Copy the pivot table to a new worksheet. Here, I copied the table to the cells B3:H4 of the new work-sheet. For example, B3 gives O11, B4 gives O21, etc. The Grand totals give you row totals (I3, I4), column totals (B5:H5), and sample size (I5).

	A	B	C	D	E	F	G	H	I
1	Count of ID	X8c							
2	X1		1	2	3	4	5	6	7 Grand Total
3		0	3	4	4	14	17	18	15 75
4		1	7	11	13	19	16	6	1 73
5	Grand Total		10	15	17	33	33	24	16 148
6									

In cell B7, type $= (I3*B5)/I5$ and hit enter. This gives you E11 in cell I7.

	A	B	C	D	E	F	G	H	I
1	Count of ID	X8c							
2	X1	1	2	3	4	5	6	7	Grand Total
3	0	3	4	4	14	17	18	15	75
4	1	7	11	13	19	16	6	1	73
5	Grand Total	10	15	17	33	33	24	16	148
6									
7		$= (I3*B5)/I5$							
8									

	A	B	C	D	E	F	G	H	I
1	Count of ID	X8c							
2	X1	1	2	3	4	5	6	7	Grand Total
3	0	3	4	4	14	17	18	15	75
4	1	7	11	13	19	16	6	1	73
5	Grand Total	10	15	17	33	33	24	16	148
6									
7		5.067568							

Click on cell B7. In the formula bar, mark I3 and click the F4 key. Similarly, mark I5 and click the F4 key. Then hit enter. If you now copy B7 and paste to cells C7:H7, B5 will change to C5:H5, but I3 (total of row 1) and I5 (sample size) will remain same. You have now the expected frequencies for the first row in cells B7:H7.

Similarly, type $(I4*B5)/I5$ in cell B8 and hit enter. This gives you E21.

Click on cell B8. In the formula bar, mark I4 and click F4 key, mark I5 and click F4 key, then hit enter. Copy B8 to C8:H8. This gives you the second row of expected frequencies.

	A	B	C	D	E	F	G	H	I
1	Count of ID	X8c							
2	X1	1	2	3	4	5	6	7	Grand Total
3	0	3	4	4	14	17	18	15	75
4	1	7	11	13	19	16	6	1	73
5	Grand Total	10	15	17	33	33	24	16	148
6									
7		5.067568	7.601351	8.614865	16.72297	16.72297	12.16216	8.108108	
8		$= ($I$4*B5)/$I5							

	A	B	C	D	E	F	G	H	I
1	Count of ID	X8c							
2	X1	1	2	3	4	5	6	7	Grand Total
3	0	3	4	4	14	17	18	15	75
4	1	7	11	13	19	16	6	1	73
5	Grand Total	10	15	17	33	33	24	16	148
6									
7		5.067568	7.601351	8.614865	16.72297	16.72297	12.16216	8.108108	
8		4.932432	7.398649	8.385135	16.27703	16.27703	11.83784	7.891892	

You now have the observed frequencies in cells B3:H4 and expected frequencies in cells B7:H8. In any blank cell, type = chitest(B3:H4,B7:H8) and hit enter. (You can type = chitest(and then mark the cells also.) Excel returns with the P value. Here the P value is 7.093E-05, which means 7.093×10^{-5}

	A	B	C	D	E	F	G	H	I
1	Count of ID	X8c							
2	X1	1	2	3	4	5	6	7	Grand Total
3	0	3	4	4	14	17	18	15	75
4	1	7	11	13	19	16	6	1	73
5	Grand Total	10	15	17	33	33	24	16	148
6									
7		5.067568	7.601351	8.614865	16.72297	16.72297	12.16216	8.108108	
8		4.932432	7.398649	8.385135	16.27703	16.27703	11.83784	7.891892	
9									
10									
11	=chitest(B3:H4,B7:H8)								

	A	B	C	D	E	F	G	H	I
1	Count of ID	X8c							
2	X1	1	2	3	4	5	6	7	Grand Total
3	0	3	4	4	14	17	18	15	75
4	1	7	11	13	19	16	6	1	73
5	Grand Total	10	15	17	33	33	24	16	148
6									
7		5.067568	7.601351	8.614865	16.72297	16.72297	12.16216	8.108108	
8		4.932432	7.398649	8.385135	16.27703	16.27703	11.83784	7.891892	
9									
10									
11	7.093E-05								

Since P value = .00007093 < .01, you can reject the null hypothesis of no relationship between X1 and X8c at a 99% level of confidence.