Biostatistics

Week #16 6/7/2022

Ch 15 – Contingency Tables

Outline

- When working with ***nominal*** data (or ***categorical*** data) that have been grouped into categories, we often arrange the counts in a tabulated format known as ***contingency table***.

Frequency Distribution

Cells contain: -Column percent -N of cases		**V007**							
		1 Strong Democrat	2 Weak Democrat	3 Independent Democrat	4 Independent	5 Independent Republican	6 Weak Republican	7 Strong Republican	*ROW TOTAL*
V002	1: Bush	**2.6** 4	**14.9** 17	**11.7** 15	**40.2** 18	**79.5** 69	**89.6** 104	**97.0** 164	*49.1* *389*
	2: Kerry	**97.4** 136	**85.1** 95	**83.1** 104	**57.6** 25	**14.2** 12	**10.4** 12	**3.0** 5	*49.2* *390*
	3: Other	**.0** 0	**.0** 0	**5.2** 7	**2.3** 1	**6.4** 6	**.0** 0	**.0** 0	*1.7* *13*
	COL TOTAL	*100.0* *140*	*100.0* *111*	*100.0* *126*	*100.0* *44*	*100.0* *87*	*100.0* *116*	*100.0* *169*	*100.0* *792*

- In the simplest case, **two *dichotomous* random variables** are involved; the rows of the table represent the outcomes of one variable, and the columns represent the outcomes of the other one.
- A contingency table is often referred as a ***two-way frequency table*** too.

Testing a contingency table

- Hypothesis testing can test a table to see whether a row variable is ***independent*** of its column variable.
- H_0 assumes that column and row outcomes are ***independent***.
- A test statistic called χ^2 (read as ***kai-square***, and spelled as ***Chi-square***) is computed, which is a random variable having its own probability density function.

Example #1

- 100 individuals are randomly sampled from a very large population.
  - Male vs female
  - Right-handed vs left-handed.
- Here the two dichotomous random variables are "***gender***" (taking two values "male" and "female") and "***handedness***" (taking two values "left-handed" and "right-handed")

	Right-handed	Left-handed
Males	**43**	**9**
Females	**44**	**4**

THEN?

Cont'd

- Usually we are interested in knowing whether there is a **correlation** between gender and handedness (left-handed or right-handed). **That is, are men more left-handed (or right-handed) than women?**
- It may certainly look true from the numbers shown below. Is it **statistically** sound?

	Right-handed	Left-handed	TOTALS
Males	43 (82.7%)	**9 (17.3%)**	52
Females	44 (91.7%)	**4 (8.3%)**	48
TOTALS	87 (87.0%)	13 (13.0%)	100

Cont'd

- **The significance of the difference** between the two proportions can be assessed (or **the association** being measured) with a variety of statistical tests including Pearson's **chi-square test**, the *G*-test, Fisher's exact test, and Barnard's test, provided the entries in the table represent individuals **randomly sampled** from the population about which we want to draw a conclusion.

Example #2

- Consider the following 2 by 2 table displaying the study investigating the effectiveness of bicycle safety helmets in preventing head injuries.

Head Injury	Wearing Helmet	
	Yes	No
Yes	17	218
No	130	428

Cont'd

- To examine the effectiveness of helmet wearing, we test the null hypothesis at α=0.05 level of significance:
  - H_0: The proportion of persons suffering head injuries for people who wore helmets at the time of accident is **the same** as the people who did not wear helmet. (有戴與沒戴, 在意外發生時, 都一樣會受傷)
  - H_A: There is a **difference** between wearing and not wearing helmet

"Expected" Contingency Table

- We will first **reconstruct** the "original" contingency table based on the null hypothesis. Resulting table is called an "expected" contingency table.
- That is, the proportions of individuals experiencing head injuries among those wearing helmets and those not wearing helmets are ***identical*** in this "expected" contingency table.

We begin by creating the total column. The purpose is to get the percentages of head injury or not from the total. Here we have roughly a 3:7 ratio,

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	17	218	235 (**29.6%**)
No	130	428	558 (**70.4%**)

We next create the total row, and know that we have 147 people wearing helmet and 646 did not. Here we know the total number of subjects are 793 (from either the total column or total row).

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	17	218	235 (**29.6%**)
No	130	428	558 (**70.4%**)
Total	**147**	**646**	793

- If we did not know the counts in the 4 blue cells, can you fill out something into them based on the 3:7 ratio in the total column?
- What hint did I give to you?
- What would you do?
- What's in your mind when you did this?

Head Injury	Wearing Helmet		
	Yes	No	Total
Yes	?	?	235 (29.6%)
No	?	?	558 (70.4%)
Total	147	646	793

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	?	?	235 (29.6%)
No	?	?	558 (70.4%)
Total	**147**	**646**	793

- Consider the following two groups of individuals:
  - ***For 147 wearing helmets***, we'd expect:
    - 147×**29.6%**=43.6 get their heads injured; and 147×**70.4%**=103.4 not injured.
  - ***For 646 not wearing helmets***, we'd expect:
    - 646×**29.6%**=191.4 get their heads injured; and 646×**70.4%**=454.6 not injured.

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	**17**	**218**	235
No	**130**	**428**	558
Total	147	646	793

Original contingency table; denoted by ***O***.

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	**43.6**	**191.4**	235.0
No	**103.4**	**454.6**	558.0
Total	147.0	646.0	793.0

Expected contingency table, ***provided that the null hypothesis is true***, denoted by ***E***.

We want to know if the deviations of these 4 cells between these two tables, that is, ***O***–***E***, are too large to be attributed to chance alone.

Chi-Square (χ^2) Test

- The chi-square test compares the observed frequencies (counts) in each category of the contingency table with the expected frequencies *given that the null hypothesis is true*.
- It is denoted by the following formula, where rc is the number of cells in the table.

$$X^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$

Cont'd

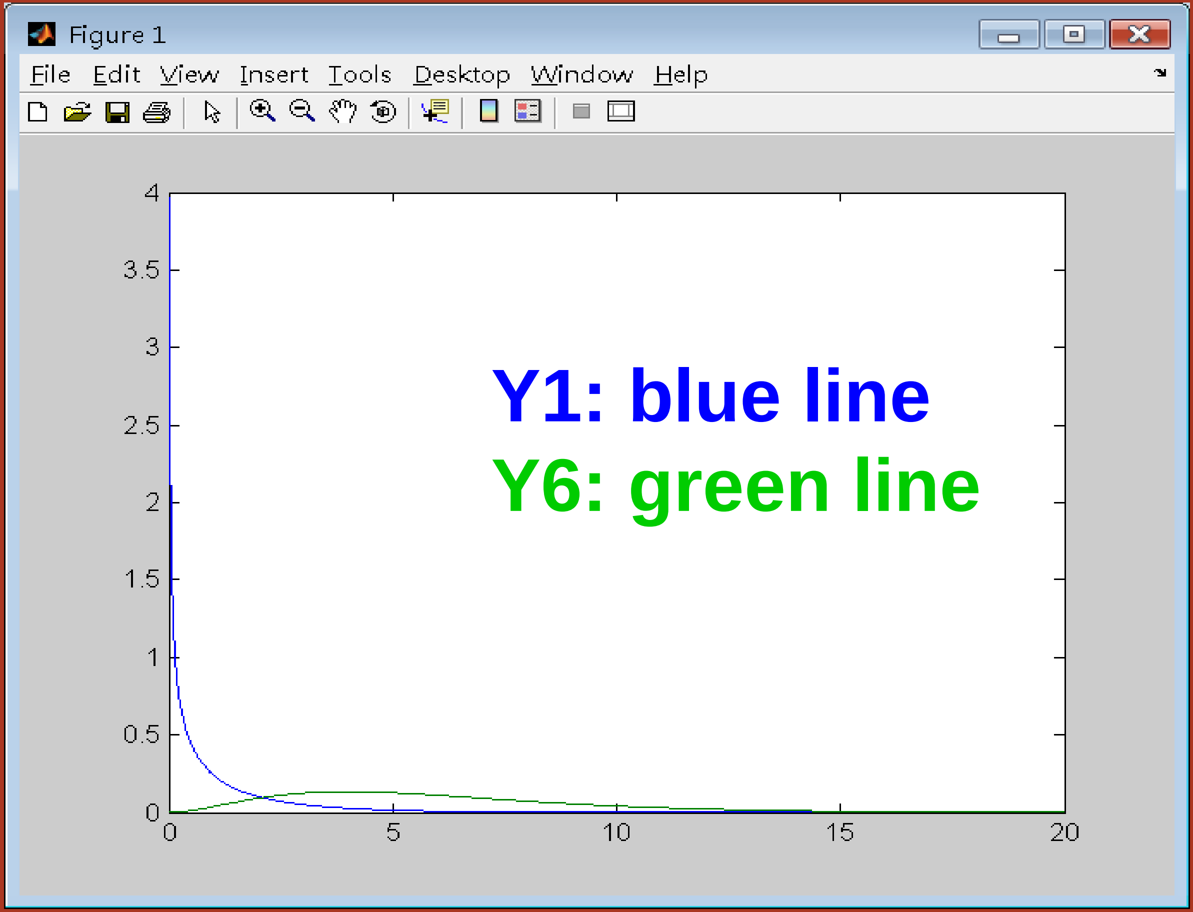
- The probability distribution of this sum is approximated by ***a chi-square (χ^2) distribution with (r–1)(c–1) degrees of freedom***, with a table of *r* rows and *c* columns.
- So a 2 by 2 table will have df= (2–1)(2–1)=1, and 3 by 4 table will have df= (3–1)(4–1)=6.
- A chi-square distribution is not symmetric.
- ***The test is one-tailed***.

See the resemblance to F-test?

>> help **chi2pdf**

CHI2PDF Chi-square probability density function (pdf). Y = CHI2PDF(X,V) returns the chi-square pdf with V degrees of freedom at the values in X.

```
>> x=0:0.01:20;
>> y1=chi2pdf(x,1);
>> y6=chi2pdf(x,6);
>> plot(x,y1,x,y6)
>>
```



>> help **chi2inv**

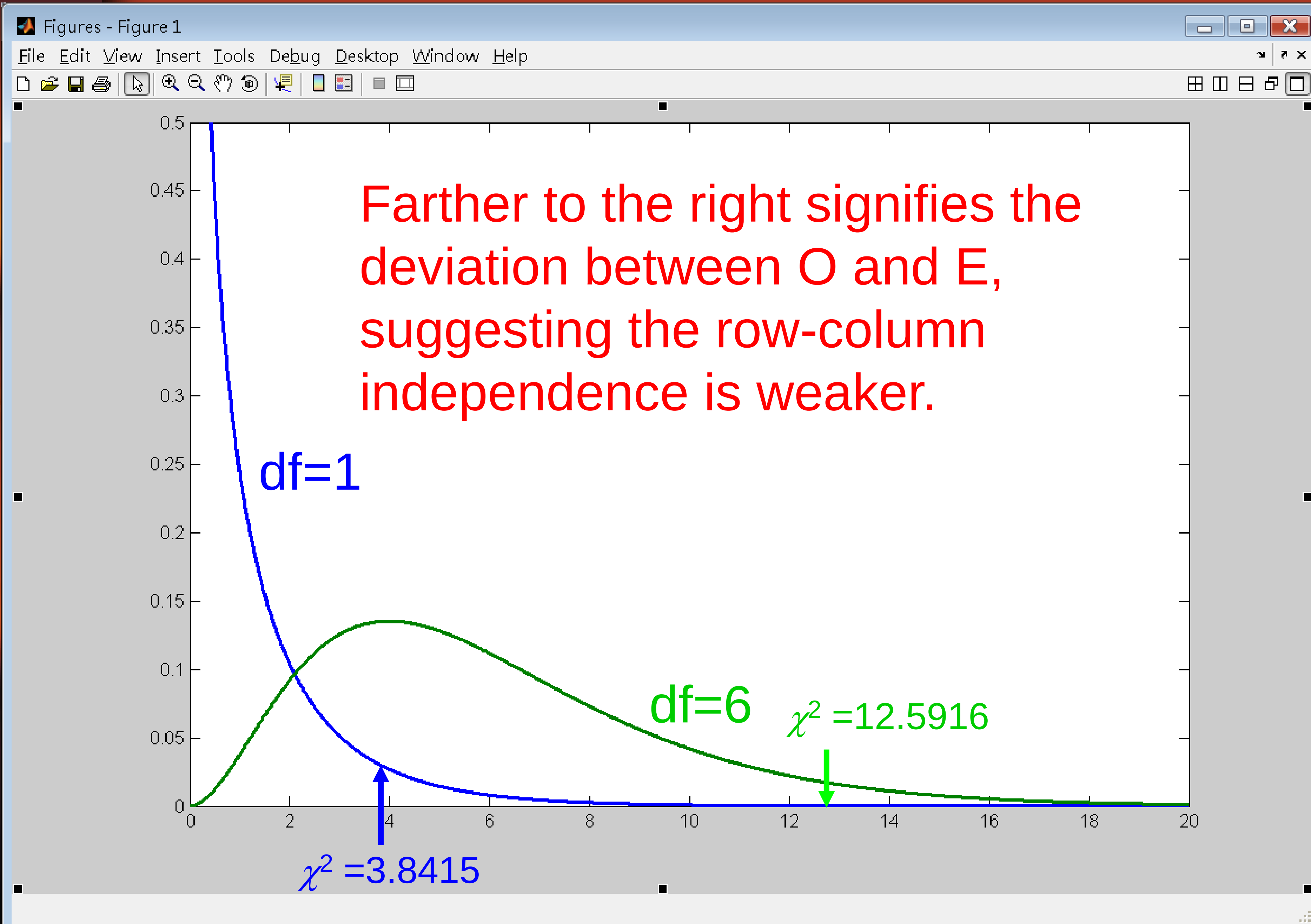
CHI2INV Inverse of the chi-square **cumulative** distribution function (cdf).

X = CHI2INV(P,V) returns the inverse of the chi-square cdf with V degrees of freedom at the values in P.

```
>> chi2inv(0.95,1)
ans =
   3.8415
>> chi2inv(0.95,6)
ans =
  12.5916
>>
```

χ^2 =3.8415 cuts off right-handed 5% off the df=1 curve.

χ^2 =12.5916 cuts off the right-handed 5% off the df=6 curve.


Figures - Figure 1
File Edit View Insert Tools Debug Desktop Window Help
Farther to the right signifies the deviation between O and E, suggesting the row-column independence is weaker.
df=1
df=6
χ^2 =12.5916
χ^2 =3.8415
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 2 4 6 8 10 12 14 16 18 20

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	17	218	235
No	130	428	558
Total	147	646	793

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	43.6	191.4	235.0
No	103.4	454.6	558.0
Total	147.0	646.0	793.0

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i} = \frac{(17-43.6)^2}{43.6} + \frac{(218-191.4)^2}{191.4}$$

$$+ \frac{(130-103.4)^2}{103.4} + \frac{(428-454.6)^2}{454.6}$$

$$= 28.3246$$

- We see that χ^2 =28.3246 is far to the right of χ^2 =3.8415 that cuts off 5% off the df=1 curve. We thus ***reject*** the null hypothesis at α=0.05.
- That is, there indeed exists a significant difference of wearing helmet or not regarding head injuries.
- p-value = 1.0258e-007

```
>> 1-chi2cdf(28.3246,1)
ans =  1.0258e-007
>>
```

Can we view by “rows” rather than “columns”?

Head Injury	Wearing Helmet		
	Yes	No	Total
Yes	17	218	235
No	130	428	558
Total	**147(18.5%)**	**646(81.5%)**	793

For 793 people, 18.5% wore helmet and 81.5% (or the remaining people out of the aforementioned 18.5%) did not.

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	?	?	235
No	?	?	558
Total	**147** **(18.5%)**	**646** **(81.5%)**	793

- Consider the following two groups of individuals :
  - ***For 235 having head injury***, we'd expect:
    - $235 \times$ **18.5%** $= 43.5$ wore helmet; and $235 \times$ **81.5%** $= 191.5$ did not.
  - ***For 558 having no head injury***, we'd expect:
    - $558 \times$ **18.5%** $= 103.2$ get their heads injured; and $558 \times$ **81.5%** $= 454.8$ not injured.

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	**43.6**	**191.4**	235.0
No	**103.4**	**454.6**	558.0
Total	147.0	646.0	793.0

Expected contingency table computed by proportion from “Total” column.

Head Injury	Wearing Helmet		Total
	Yes	No	
Yes	**43.5**	**191.5**	235.0
No	**103.2**	**454.8**	558.0
Total	147.0	646.0	793.0

Expected contingency table computed by proportion from “Total” row.

We can see the two expected tables “the same” except some round-off digits.

Example #3

- Instead of 2 by 2 table we saw earlier, we now have a 2 by 3 table to consider.
- Results from 575 autopsies (解剖驗屍) were compared to the cause of death listed on the certificates (死亡證明).

Hospital	Death Certificate Status			Total
	Confirmed. Accurate.	Inaccurate. No change.	Incorrect. Recoding.	
A (Community H)	**157**	**18**	**54**	229
B (University H)	**268**	**44**	**34**	346
Total	425	62	88	575

Cont'd

- We would like to determine whether the results of this study suggest **different practices** in completing death certificates at the two hospitals.
- The null hypothesis could be either of the following two (at significance level 0.05):
  - H_0: **within each category of certificate status**, the proportions of death certificates in hospital A are identical.
  - H_0: there is no association between hospital and death certificate status.

Building E based on row-proportions of total.

Hospital	Death Certificate Status			Total
	Confirmed. Accurate.	Inaccurate. No change.	Incorrect. Recoding.	
A (Community H)	157	18	54	229 (39.83%)
B (University H)	268	44	34	346 (60.17%)
Total	425	62	88	575

We would have the expected counts as:

Hospital	Death Certificate Status			Total
	Confirmed. Accurate.	Inaccurate. No change.	Incorrect. Recoding.	
A (Community H)	**169.3** (425x39.83%)	**24.7**	**35.0**	229
B (University H)	**255.7**	**37.3**	**53.0**	346
Total	425.0	62.0	88.0	575

Building E based on column-proportions of total.

Hospital	Death Certificate Status			Total
	Confirmed. Accurate.	Inaccurate. No change.	Incorrect. Recoding.	
A (Community H)	157	18	54	229
B (University H)	268	44	34	346
Total	425 (73.91%)	62(10.78%)	88(15.31%)	575

We could also have the expected counts as:

Hospital	Death Certificate Status			Total
	Confirmed. Accurate.	Inaccurate. No change.	Incorrect. Recoding.	
A (Community H)	**169.3** (73.91%x229)	**24.7**	**35.0**	229
B (University H)	**255.7**	**37.3**	**53.0**	346
Total	425.0	62.0	88.0	575

Conclusion

- The chi-square can be computed as **21.62**.
- This allows us to reject the null hypothesis (p-value = 1-chi2cdf(21.62,2) = 2.0197e-005)
- That is, hospital A and hospital B are indeed different.
- For example, it is clear (from the original table) that hospital A apparently issued more incorrect certificates that required recoding, suggesting that a community hospital requires improving its practice in issuing a death certificate.

Hospital	Death Certificate Status			Total
	Confirmed. Accurate.	Inaccurate. No change.	Incorrect. Recoding.	
A (Community H)	157	18	**54**	229
B (University H)	268	44	34	346
Total	425	62	88	575

In-class practice #1

- Is left-handedness related to gender, at a level of significance set to 0.05?

	Right-handed	Left-handed
Males	43	9
Females	44	4

Original table:

	Right-handed	Left-handed	TOTALS
Males	43 (82.7%)	9 (17.3%)	52
Females	44 (91.7%)	4 (8.3%)	48
TOTALS	87 (87.0%)	13 (13.0%))	100

Expected table:

	Right-handed	Left-handed	TOTALS
Males			52
Females			48
TOTALS	87 (87.0%)	13 (13.0%))	100

Chi-square = ?

P-value = ?

Conclusion = ?

Original table:

	Right-handed	Left-handed	TOTALS
Males	43 (82.7%)	9 (17.3%)	52
Females	44 (91.7%)	4 (8.3%)	48
TOTALS	87 (87.0%)	13 (13.0%)	100

Expected table:

	Right-handed	Left-handed	TOTALS
Males	E1=87*52/100	E2=13*52/100	52
Females	E3=87-E1	E4=13-E2	48
TOTALS	87 (87.0%)	13 (13.0%)	100

Chi-square = ?

P-value = ?

Conclusion = ?

Original table:

	Right-handed	Left-handed	TOTALS
Males	43 (82.7%)	9 (17.3%)	52
Females	44 (91.7%)	4 (8.3%)	48
TOTALS	87 (87.0%)	13 (13.0%))	100

Expected table:

	Right-handed	Left-handed	TOTALS
Males	E1=45.24	E2=6.76	52
Females	E3=41.76	E4=6.24	48
TOTALS	87 (87.0%)	13 (13.0%)	100

Chi-square = ?

P-value = ?

Conclusion = ?

```
>> O=[43 9 44 4];
>> E=[45.24 6.76 41.76 6.24];
>> (O-E).^2./E
ans =    0.1109    0.7422    0.1202    0.8041
>> X2=sum(ans)
X2 = 1.7774

>> 1-chi2cdf(X2,1)
ans =  0.1825
>>
```

Chi-square statistic = 1.7774. The P-value is 0.1825, which greater than 0.05. We thus do not reject the null hypothesis, which claims no association between gender and handedness.

In-class practice #2

- An outbreak of gastroenteritis – an inflammation of the membranes of the stomach and small intestine, was recorded following a lunch served in the cafeteria.
- Is having sandwich a cause for sickness, at level significance set to 0.05?

	Had sandwich	
	Yes	No
Sick	**109**	**4**
Not sick	**116**	**34**

	Had sandwich		Total
	Yes	No	
Sick	**109**	**4**	113
Not sick	**116**	**34**	150
Total	225	38	263

Original table

	Had sandwich		Total
	Yes	No	
Sick			113
Not sick			150
Total	225	38	263

Expected table

Chi-square = ?

P-value = ?

Conclusion?

	Had sandwich		Total
	Yes	No	
Sick	**109**	**4**	113
Not sick	**116**	**34**	150
Total	225	38	263

	Had sandwich		Total
	Yes	No	
Sick	E1=225*113/263	E2=113-E1	113
Not sick	E3=225-E1	E4=38-E2	150
Total	225	38	263

Chi-square = ?

P-value = ?

Conclusion?

	Had sandwich		Total
	Yes	No	
Sick	**109**	**4**	113
Not sick	**116**	**34**	150
Total	225	38	263

	Had sandwich		Total
	Yes	No	
Sick	E1=96.6730	E2=16.3270	113
Not sick	E3=128.3270	E4=21.6730	150
Total	225	38	263

Chi-square = ?

P-value = ?

Conclusion?

```
>> O=[109 4 116 34];
>> E=[96.673 16.327 128.327 21.673];
>> (O-E).^2./E
ans =    1.5718    9.3070    1.1841    7.0113
>> X2=sum(ans)
X2 =   19.0742
>> 1-chi2cdf(X2,1)
ans =  1.2573e-005
>>
```

Chi-square statistic = 19.0724. The p-value is small enough ($p < 0.001$) to reject the null hypothesis. We then link the sickness with having sandwich.