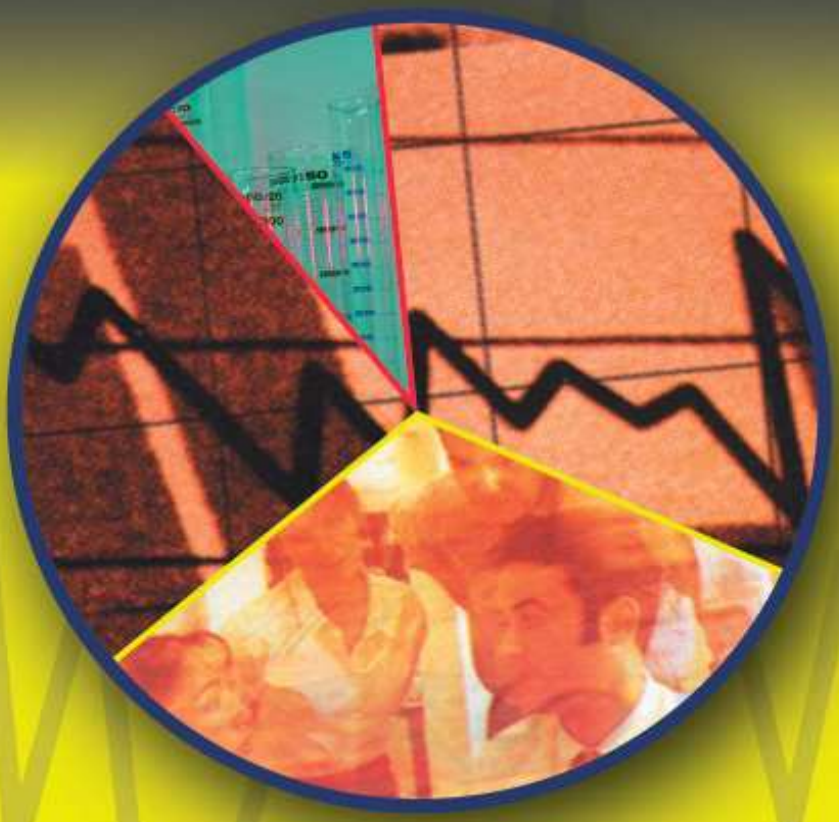
# Statistics Made Simple

## *for* Healthcare and Social Science Professionals and Students



Wong Kam Cheong • Phua Kai Lit

# Preface

We decided to write this book because we noticed that many people find statistics to be confusing and mystifying. They also find standard textbooks on statistics to be difficult to understand and to use. We are optimists and believe that it is not difficult to learn basic statistics and apply statistical tests to analyze data and test hypotheses. This can be done if the use of mathematical and statistical notation and jargon are minimized and if attempts are made to explain things as simply and clearly as possible for those who want to use statistics as a tool rather than ponder over the intricacies of statistical theory. The authors believe that this book will help you to understand, appreciate and apply statistics (happily!) in your work and studies.

The fundamentals of statistics, methods of sampling, types of data, levels of measurement, measures of central tendency, measures of dispersion, and hypothesis testing are covered in Chapters 1, 2 and 3. Chapter 4 deals with $\chi^2$ (chi-square) test that can be applied for nominal (categorical) and ordinal data. Categorical data that are commonly collected in studies to investigate relative risk and odds ratio are dealt with in Chapter 5. Statistical tools for handling and analyzing interval data are elaborated in Chapter 6 (t-test and Z-test), Chapter 7 (Analysis of Variance / ANOVA, comparing averages of two and more sources) and Chapter 8 (Correlation, Regression Analysis). The steps to be taken in analyzing quantitative data are summarized in Appendix I. Information about some useful software packages, references and website links are provided in Appendix II. Readers can resort to statistical tables or Microsoft Excel software to obtain certain statistical indices and values. The relevant Microsoft Excel commands are summarized in Appendix III. Some residual plots related to regression analysis are presented in Appendix IV. The statistical tables for $\chi^2$ (Chi-square) distribution, t-distribution, Standard Normal distribution, and F-distribution are provided in Appendix V, VI, VII, and VIII respectively.

Dr Wong is a statistician trained in a Master of Science (statistics) program at the National University of Singapore. He received his Bachelor of Medicine and Bachelor of Surgery (MBBS) from the University of Queensland, Australia. Dr Phua is a social scientist trained in a PhD program (Johns Hopkins University, United States of America) which emphasizes the acquisition of statistical skills. Dr Phua is currently a senior lecturer in Community Medicine at the International Medical University in Kuala Lumpur, Malaysia. Both authors like statistics and believe that it is a useful and elegant tool for making sense of data. Welcome to the world of statistics!

Dr. Wong Kam Cheong
*MBBS, BE, MSc.*

Dr. Phua Kai Lit
*BA, PhD.*

March 2006 http://www.btm.upm.edu.my/penerbit/

## Authors' Background

Dr. Wong Kam Cheong received his Bachelor of Medicine and Bachelor of Surgery (MBBS) from the University of Queensland (Australia), and Bachelor of Engineering (Mechanical) and Master of Sciences (Statistics) from the National University of Singapore. He has worked as a statistician and engineer in Singapore and Malaysia, and as a medical research associate in the Centre for Chronic Disease at the University of Queensland (Australia). He is currently working as a medical doctor in Australia. His past awards include a research scholarship from the Dartmouth Medical School (Dartmouth College in the United States of America) which is one of the Ivy Leagues universities.

Dr. Phua Kai Lit received his BA in Public Health and Population Studies from the University of Rochester (United States of America) and PhD in Sociology (Medical Sociology) from Johns Hopkins University (United States of America). He has worked as a research statistician with the Maryland Department of Health and Mental Hygiene in Baltimore, USA. Dr. Phua is currently a senior lecturer with the Community Medicine section of the International Medical University in Kuala Lumpur, Malaysia. His past awards include an Asian Public Intellectual Senior Fellowship from the Nippon Foundation.

**The following is a chapter extracted from the book**

# "Statistics Made Simple for Healthcare and Social Science Professionals and Students"

**Authors: Dr. Wong Kam Cheong, Dr. Phua Kai Lit.**

**International Standard Book Number (ISBN) = 983-3455-03-4.**

# Chapter 4

- Introduction to the $\chi^2$ Test (Chi-square Test)
- Assumptions of the $\chi^2$ Test
- Using the $\chi^2$ Test (Worked Example)

### 4.1 Introduction to the $\chi^2$ test (Chi-square Test)

In this chapter, you will be introduced to a very useful test called the $\chi^2$ test (also known as the "chi-square test of association" or "chi-square test of independence"). It is used to determine if there is a statistically significant association between two variables measured at the nominal level. It can also be used to test for association between two variables measured at the ordinal level.

### 4.2 Assumptions of the $\chi^2$ Test:

- The data should be nominal data or ordinal data
- The samples should be random samples
- The 'expected value' of each cell should be at least 5 (if not, the categories can be combined to overcome this problem). The 'expected value' will be explained later on in this chapter.

### 4.3 Using the $\chi^2$ Test (Worked Example)

Here is an example to illustrate how the $\chi^2$ test is used. Suppose you wish to test if there is an association between gender ("sex") and anorexia nervosa in the case of Malaysian teenagers. Anorexia nervosa is an eating disorder characterized by refusal to maintain a normal minimal body weight. Sufferers believe that they are overweight and will refuse to eat properly even if they are actually seriously underweight. You could select a random sample of 100 male Malaysian teenagers from the population of all male Malaysian teenagers and a second random

sample of 100 female Malaysian teenagers from the population of all female Malaysian teenagers. Next you would check how many of the female teenagers suffer from anorexia nervosa vis-à-vis how many of the male teenagers suffer from anorexia nervosa. Then you would use the $\chi^2$ test to see if the difference in prevalence of anorexia nervosa between the males and the females is statistically significant or due to chance alone. The following denotations are used when we refer to a 'cell' in a table prepared for the $\chi^2$ test:

| | |
|---|---|
| "Cell 1-1" | "Cell 1-2" |
| "Cell 2-1" | "Cell 2-2" |

The steps involved in analyzing data using the $\chi^2$ test are as follows:

[Step 1] Place the data in a 2 X 2 "contingency table" (two by two contingency table) of OBSERVED VALUES (Table 4.1).

**Table 4.1**: Gender and occurrence of anorexia nervosa, sample of 100 male Malaysian teenagers and sample of 100 female Malaysian teenagers (OBSERVED VALUES)

| | | Anorexia Nervosa | |
|---|---|---|---|
| | | Yes | No |
| Gender | Male | 3 | 97 |
| | Female | 10 | 90 |

If there is a relationship between gender and anorexia nervosa, e.g. *if females are more likely to suffer from the disease than males*, then the number in cell 1-1 (top left hand corner cell) would be small and the number in cell 2-1 (bottom left hand corner cell) would be large (Table 4.2).

**Table 4.2**: Female teenagers are more likely to suffer from anorexia nervosa

| | | Anorexia Nervosa | |
|---|---|---|---|
| | | Yes | No |
| Gender | Male | Small<br>1 | 99 |
| | Female | Large<br>11 | 89 |

If there is a relationship between gender and anorexia nervosa in the other direction, i.e. *if males are more likely to suffer from the disease than females*, then the number in cell 1-1 (top left hand corner cell) would be large and the number in cell 2-1 (bottom left hand corner cell) would be small (See Table 4.3).

**Table 4.3:** Male teenagers are more likely to suffer from anorexia nervosa

| | | Anorexia Nervosa | |
|---|---|---|---|
| | | Yes | No |
| Gender | Male | Large<br>13 | 87 |
| | Female | Small<br>2 | 98 |

If there is NO relationship between gender and anorexia nervosa, e.g. *if both females and males are equally likely to suffer from the disease*, then the number in cell 1-1 (top left hand corner cell) would be large and the number in cell 2-1 (bottom left hand corner cell) would also be large (Table 4.4).

**Table 4.4:** Both male and female teenagers are equally likely to suffer from anorexia nervosa

| | | Anorexia Nervosa | |
|---|---|---|---|
| | | Yes | No |
| Gender | Male | Large<br>12 | 88 |
| | Female | Large<br>11 | 89 |

If there is NO relationship between gender and anorexia nervosa in the other direction, i.e. *if both males and females are equally unlikely to suffer from the disease,* then the number in cell 1-1 (top left hand corner cell) would be small and the number in cell 2-1 (bottom left hand corner cell) would be small (Table 4.5).

**Table 4.5:** Both male and female teenagers are unlikely to suffer from anorexia nervosa

| | | Anorexia Nervosa | |
|---|---|---|---|
| | | Yes | No |
| Gender | Male | Small<br>3 | 97 |
| | Female | Small<br>2 | 98 |

[Step 2] After drawing the table of OBSERVED VALUES **(Table 4.1)**, we need to construct a table of EXPECTED VALUES (See Table 4.6).

**Table 4.6:** Gender and occurrence of anorexia nervosa, sample of 100 male Malaysian teenagers and sample of 100 female Malaysian teenagers (EXPECTED VALUES)

| | | Anorexia Nervosa | | |
|---|---|---|---|---|
| | | Yes | No | |
| Gender | Male | a | b | Row 1 total |
| | Female | c | d | Row 2 total |
| | | Column 1 total | Column 2 total | |

The EXPECTED VALUE of "a" in cell 1-1 would be

$$\frac{\text{(Row 1 total) X (Column 1 total)}}{n}$$

where n = sum of sample sizes, i.e.100 males + 100 females = 200

Similarly, the EXPECTED VALUE of "b" in cell 1-2 would be

$$\frac{\text{(Row 1 total) X (Column 2 total)}}{n}$$

The EXPECTED VALUE of "c" in cell 2-1 would be

$$\frac{\text{(Row 2 total) X (Column 1 total)}}{n}$$

The EXPECTED VALUE of "d" in cell 2-2 would be

$$\frac{\text{(Row 2 total) X (Column 2 total)}}{n}$$

Thus, based on the numbers given in **Table 4.1**, the EXPECTED VALUES for the entire table would be:

**Table 4.7:** Gender and occurrence of anorexia nervosa, sample of 100 male Malaysian teenagers and sample of 100 female Malaysian teenagers (EXPECTED VALUES)

| | | Anorexia Nervosa | |
|---|---|---|---|
| | | Yes | No |
| Gender | Male | $\frac{100\times13}{200}=6.5$ | $\frac{100\times187}{200}=93.5$ |
| | Female | $\frac{100\times13}{200}=6.5$ | $\frac{100\times187}{200}=93.5$ |

[Step 3] Calculate an 'index' called the $\chi^2$ for each cell and add them all up to get the cumulative $\chi^2$. The formula for the cumulative $\chi^2$ is:

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

where $O$ = observed value found in a particular cell in the table

$E$ = expected value associated with a particular cell in the table

$$\chi^2 = \frac{(3-6.5)^2}{6.5} + \frac{(97-93.5)^2}{93.5} + \frac{(10-6.5)^2}{6.5} + \frac{(90-93.5)^2}{93.5} = 4.031$$

The observed and expected values in the above equation are obtained from Table 4.1 and Table 4.7 respectively.

[Step 4] Once we get the cumulative $\chi^2$, we should compare this value to the corresponding "critical value" in a $\chi^2$ table to see if it is statistically significant. If it is statistically significant (it equals or exceeds the "critical value"), we would reject $H_0$ and accept $H_1$

$H_0$ : There is no association between gender and suffering from anorexia nervosa. Any association seen is due to chance alone.

$H_1$ : There is a statistically significant association between gender and suffering from anorexia nervosa.

To find the "critical value" of $\chi^2$ for a 2 X 2 contingency table at the 0.05 level, we would look at the "critical value" for degree of freedom = 1 and probability level of 0.05. The table (Table 4.8) shows that the critical value is 3.841 {The "degree of freedom" is equal to (number of rows minus 1) X (number of columns minus 1) = (2 – 1) X (2 – 1) = 1}.

**Table 4.8:** $\chi^2$ table (truncated and simplified)

| Degree of Freedom | Probability of 0.05 | Probability of 0.01 |
|---|---|---|
| 1 | 3.841 | 6.635 |
| 2 | 5.991 | 9.210 |
| 3 | 7.815 | 11.345 |
| 4 | 9.488 | 13.277 |
| 5 | 11.070 | 15.086 |
| 6 | 12.592 | 16.812 |

Alternatively, the critical $\chi^2$ value can be obtained by using statistical software such as Microsoft Excel. Type the following command in a field in a Microsoft Excel spreadsheet: =CHIINV(probability, degree of freedom) (Appendix III part 1a). For this example, the command is =CHIINV(0.05, 1) A value of "3.841" will be returned.

*Decision Rule of Less Stringent Test:*

Reject $H_0$ and accept $H_1$ if the calculated cumulative $\chi^2$ is greater than or equal to the critical value of 3.841 when testing at the 0.05 level with degree of freedom of 1 (where the value 0.05 refers to the probability of $H_1$ occurring by chance).

Alternatively, reject $H_0$ and accept $H_1$ if the *p*-value is less than or equal to 0.05. The *p*-value can be obtained by typing the following command in a field in a Microsoft Excel spreadsheet: =CHIDIST(computed cumulative chi-square value, degree of freedom) (Appendix III part 1b). In this example, the cumulative chi-square value is computed by adding up the $\chi^2$

(chi-square) value for each cell. This is found to be 4.031. Hence, the command for obtaining p-value from Microsoft Excel is =CHIDIST(4.031, 1). A value of "0.045" will be returned.

*Decision Rule of More Stringent Test:*

Reject $H_0$ and accept $H_1$ if the calculated cumulative $\chi^2$ is greater than or equal to the critical value of 6.635 when testing at the 0.01 level with degree of freedom of 1 (where the value 0.01 refers to the probability of $H_1$ occurring by chance). The critical value "6.635" is obtained from Table 4.8 or by Microsoft Excel command =CHIINV(0.01, 1).

Alternatively, reject $H_0$ and accept $H_1$ if the p-value is less than or equal to 0.01. Thus, in the example given in Table 4.1, the cumulative $\chi^2$ is found to be 4.031. When this cumulative chi-square value is used for the less stringent test ($\alpha$= 0.05 at degree of freedom = 1), it is found to exceed the critical value of 3.841. Also, the *p*-value (0.045) is less than 0.05. Therefore, we reject $H_0$ and accept $H_1$ and conclude that there is a statistically significant association between gender and suffering from anorexia nervosa. When we reject $H_0$ and accept $H_1$, the probability that $H_1$ has occurred by chance is less than 0.05 or less than 5%.

However, when we use the cumulative $\chi^2$ value of 4.031 for the more stringent test ($\alpha$= 0.01 at degree of freedom = 1), it is found to be less than the critical value of 6.635. Also, the *p*-value (0.045) is more than 0.01. Therefore, it has passed the less stringent test but failed the more stringent test. Hence, when we reject $H_0$ and accept $H_1$, the probability that $H_1$ has occurred by chance is less than 5% but more than 1%.

http://www.btm.upm.edu.my/penerbit/

**ORDER FORM**

| ***Quantity*** | ***Title*** | **ISBN** | ***Price*** |
|---|---|---|---|
| | Statistics Made Simple for Healthcare and Social Science Professionals and Students | 983-3455-03-4 | [ ] x US$6 = [ ]<br>Or<br>[ ] x RM20 = [ ] |
| | | | |
| | | | |
| | | | |
| | | | |
| | ***Total*** | | US$<br>Or<br>RM |

I hereby enclose a cheque/bank draft/money order No.: _____ for the sum of US$ / RM* _____ payable to **BENDAHARI UNIVERSITI PUTRA MALAYSIA** for the above order.

Name: _____

Institute:_____

(capital letter)

Address:

_____

_____

Postcode:_____ State : _____

Country : _____

Tel. No: _____Fax No. _____E-mail : _____

*Circle either US$ or RM. (Please contact the following personnel for enquiry about postage cost.)

Send your order to: **UPM Press, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.**

Tel. No.: 603 8946 8851 / 8855 / 8854. Fax No.: 603 8941 6172.

E-mail: *penerbit@putra.upm.edu.my*