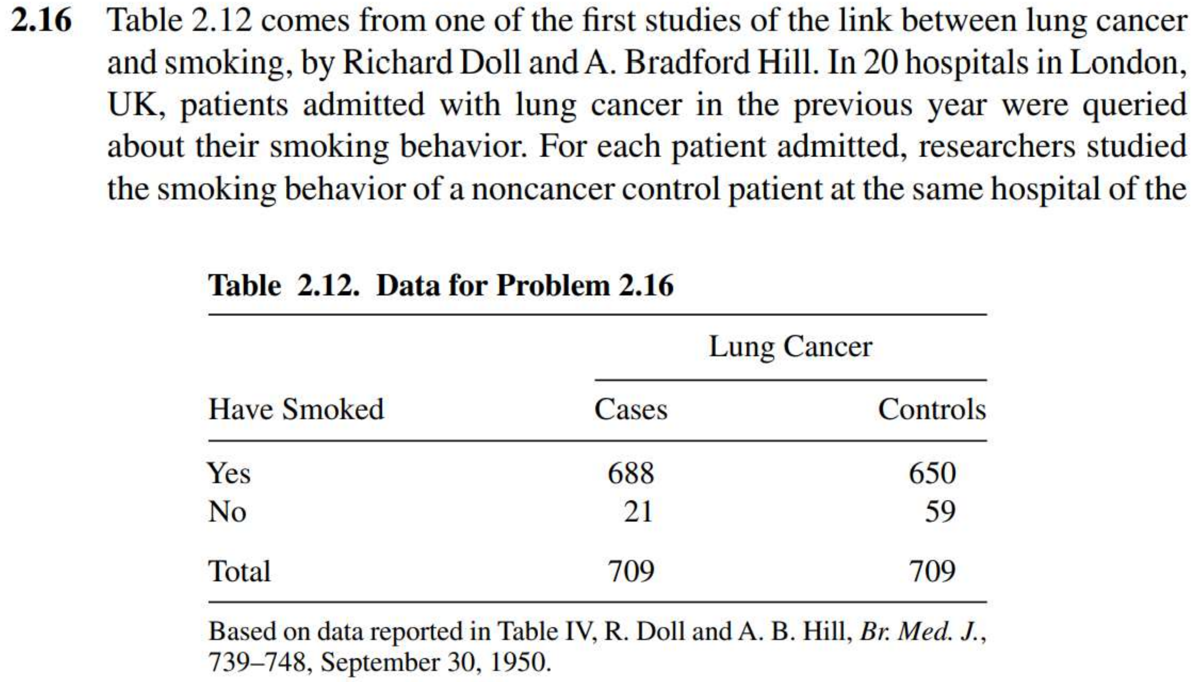
第七題: Problem 2.16

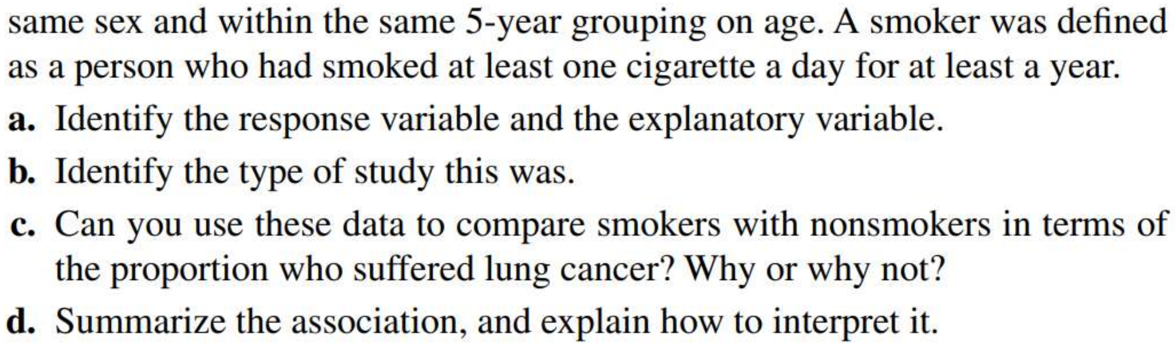


2.16 Table 2.12 comes from one of the first studies of the link between lung cancer and smoking, by Richard Doll and A. Bradford Hill. In 20 hospitals in London, UK, patients admitted with lung cancer in the previous year were queried about their smoking behavior. For each patient admitted, researchers studied the smoking behavior of a noncancer control patient at the same hospital of the

Table 2.12. Data for Problem 2.16

	Lung Cancer	
Have Smoked	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Based on data reported in Table IV, R. Doll and A. B. Hill, *Br. Med. J.*, 739–748, September 30, 1950.



same sex and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

a. Identify the response variable and the explanatory variable.

b. Identify the type of study this was.

c. Can you use these data to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer? Why or why not?

d. Summarize the association, and explain how to interpret it.

(a)

explanatory variable 是 Have smoked 與否。

response variable 是 cancer 罹患者的數目。

(b)

這是一個 case-control 的 study，因為有一個 control variable。

(c)

因為罹患 cancer 的病因有可能是患者本身就身體不健康。即使在 control group 的人沒有罹患 cancer，他們也不見得是健康的。並不一定是因為抽菸所引起的。

因此就 smokers 和 non-smokers 在罹患 lung cancer 的比例上，我們並不能下結論說抽煙與否和

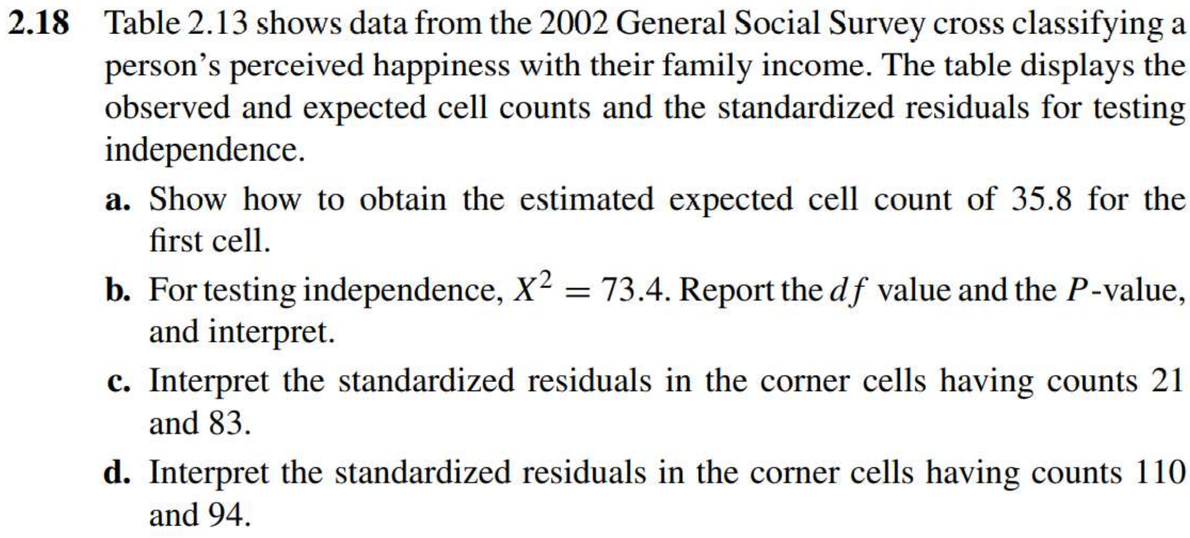
罹患 lung cancer 有絕對的關係。

(d)

Theta=(a/c)/(b/d)=ad/bc=(688x59)/(650x21)=2.9738>1

根據罹患 lung cancer 的 odds ratio， 有抽菸的人罹患 lung cancer 的頻率比沒抽菸的高 2.9738 倍。

第 8 題: Problem 2.18



2.18 Table 2.13 shows data from the 2002 General Social Survey cross classifying a person's perceived happiness with their family income. The table displays the observed and expected cell counts and the standardized residuals for testing independence.

a. Show how to obtain the estimated expected cell count of 35.8 for the first cell.

b. For testing independence, $X^2 = 73.4$. Report the df value and the P-value, and interpret.

c. Interpret the standardized residuals in the corner cells having counts 21 and 83.

d. Interpret the standardized residuals in the corner cells having counts 110 and 94.

(e) Since both variables are ordinal, test the association between them by using

(i) equally spaced scores for both variables

(ii) midranks for both variables

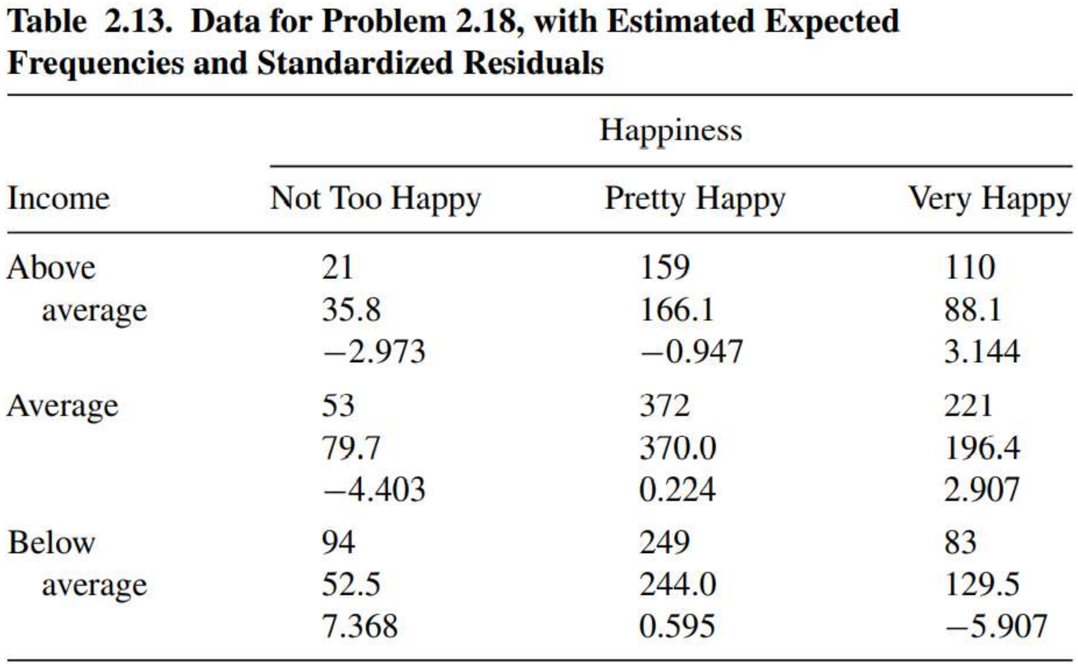


Table 2.13. Data for Problem 2.18, with Estimated Expected Frequencies and Standardized Residuals

	Happiness		
Income	Not Too Happy	Pretty Happy	Very Happy
Above average	21 35.8 −2.973	159 166.1 −0.947	110 88.1 3.144
Average	53 79.7 −4.403	372 370.0 0.224	221 196.4 2.907
Below average	94 52.5 7.368	249 244.0 0.595	83 129.5 −5.907

(a)

首先，我們先算出 first row 的 sum，為(21+159+110) =290。再來我們去算 first column 的 sum，為(21+53+94) =168。表中的 overall sum 為(21+53+94+159+372+249+110+221+83)=1362。所以所求=(290x168)/1362=35.77=35.8。

(b)

The degree of freedom for x^2 is (3-1)x(3-1)=4

我們用 chi-square table 查表看 X^2=73.4 和 df=4 發現 P value 趨近於 0 => This P-value is less than usual level of significance (say 0.05)。這代表 family income and happiness are significantly associated with each other，effect of happiness on family income is very unlikely to have arisen purely by chance.

(c)

最左上角有著 21 counts 的那個 cell，其 standardized residual 為-2.973。=> there is extremely strong evidence that there were fewer “not too happy” families with above average income than the hypothesis of independence predicts.

最右下角有著 83 counts 的那個 cell，其 standardized residual 為-5.907。=> there is extremely strong evidence that there were very fewer “very happy” families with below average income than the hypothesis of independence predicts.

(d)

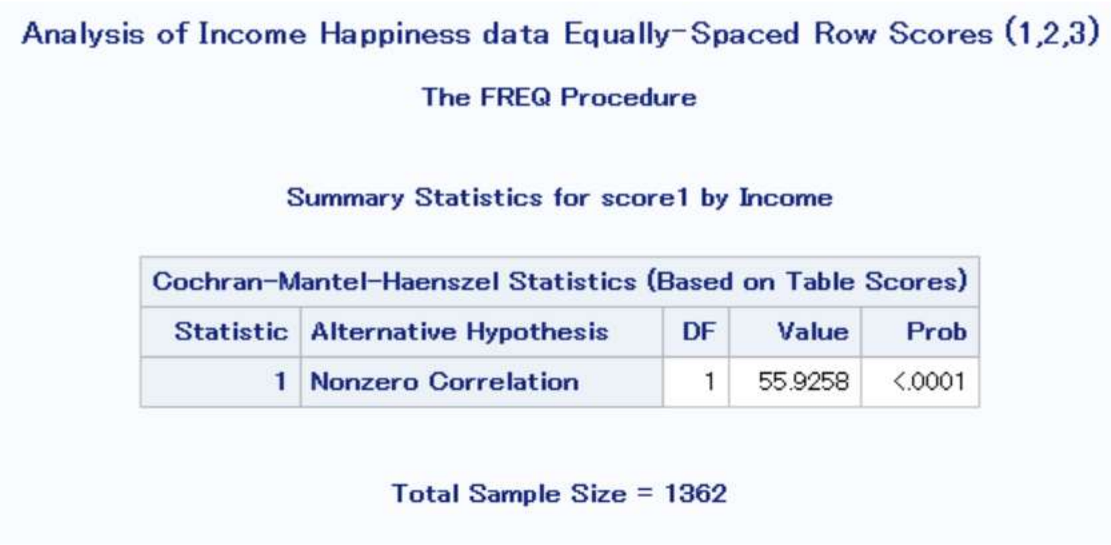
最右上角有著 110 counts 的那個 cell，其 standardized residual 為 3.144。=> there is extremely strong evidence that there are more “very happy” families with above average income than the hypothesis of independence predicts.

最左下角有著 94 counts 的那個 cell，其 standardized residual 為 7.368。=> there is extremely strong evidence that there were much more “not too happy” families with below average income than the hypothesis of independence predicts.

(e)

(i) equally spaced scores for both variables

從下面的數據可以看出，因為 p-value<0.0001，we reject the null hypothesis. There is a strong association between income and happiness.



Analysis of Income Happiness data Equally-Spaced Row Scores (1,2,3)

The FREQ Procedure

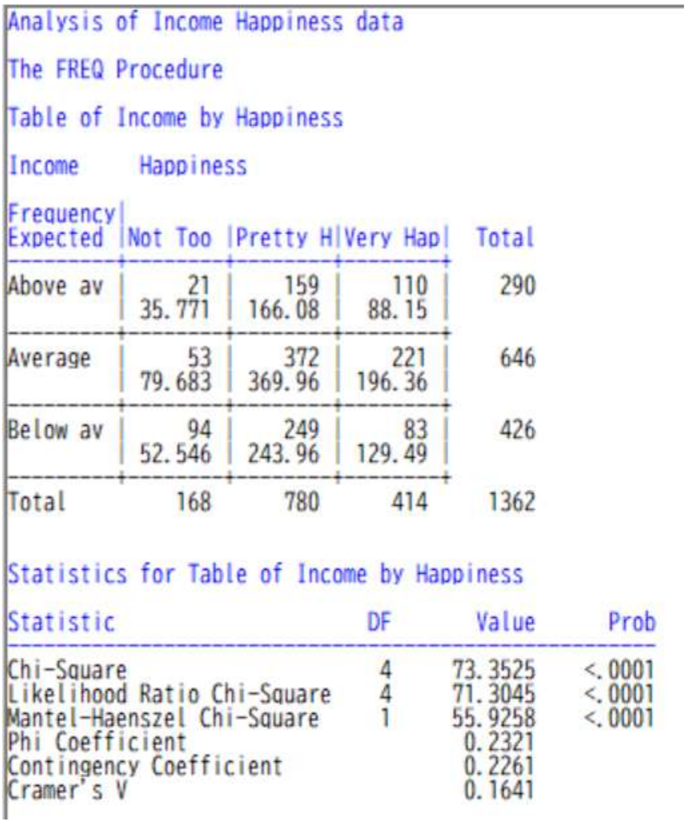
Summary Statistics for score1 by Income

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	55.9258	<.0001

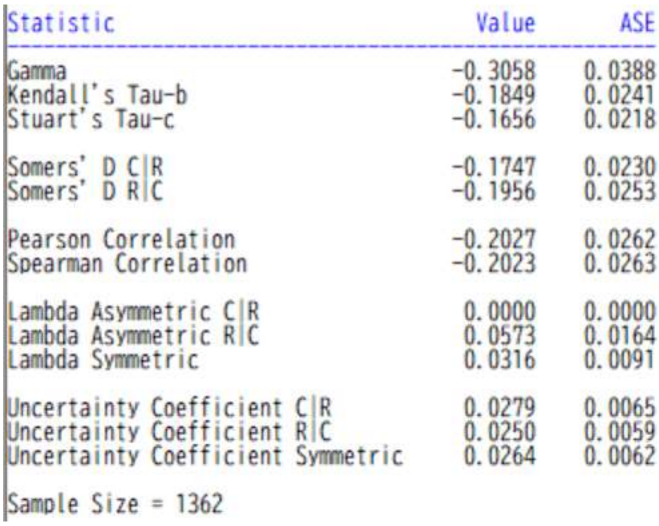
Total Sample Size = 1362

(ii) midranks for both variables

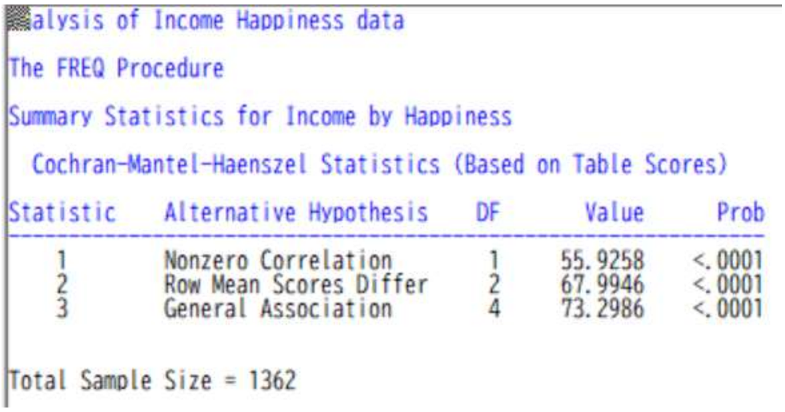
從下面的數據可以看出，因為 p-value<0.0001，we reject the null hypothesis. There is a strong association between income and happiness.



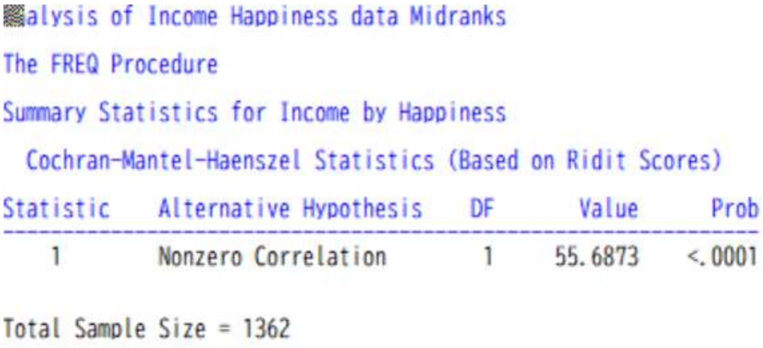
```
Analysis of Income Happiness data

The FREQ Procedure

Table of Income by Happiness

Income     Happiness

Frequency|
Expected |Not Too |Pretty H|Very Hap|  Total
---------+--------+--------+--------+
Above av |     21 |    159 |    110 |    290
         | 35.771 | 166.08 |  88.15 |
---------+--------+--------+--------+
Average  |     53 |    372 |    221 |    646
         | 79.683 | 369.96 | 196.36 |
---------+--------+--------+--------+
Below av |     94 |    249 |     83 |    426
         | 52.546 | 243.96 | 129.49 |
---------+--------+--------+--------+
Total         168      780      414     1362

Statistics for Table of Income by Happiness

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     4     73.3525    <.0001
Likelihood Ratio Chi-Square    4     71.3045    <.0001
Mantel-Haenszel Chi-Square     1     55.9258    <.0001
Phi Coefficient                       0.2321
Contingency Coefficient               0.2261
Cramer's V                            0.1641
```



```
Statistic                              Value      ASE
------------------------------------------------------
Gamma                                 -0.3058   0.0388
Kendall's Tau-b                       -0.1849   0.0241
Stuart's Tau-c                        -0.1656   0.0218

Somers' D C|R                         -0.1747   0.0230
Somers' D R|C                         -0.1956   0.0253

Pearson Correlation                   -0.2027   0.0262
Spearman Correlation                  -0.2023   0.0263

Lambda Asymmetric C|R                  0.0000   0.0000
Lambda Asymmetric R|C                  0.0573   0.0164
Lambda Symmetric                       0.0316   0.0091

Uncertainty Coefficient C|R            0.0279   0.0065
Uncertainty Coefficient R|C            0.0250   0.0059
Uncertainty Coefficient Symmetric      0.0264   0.0062

Sample Size = 1362
```



```
▒alysis of Income Happiness data

The FREQ Procedure

Summary Statistics for Income by Happiness

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
-----------------------------------------------------------------
    1        Nonzero Correlation        1     55.9258    <.0001
    2        Row Mean Scores Differ     2     67.9946    <.0001
    3        General Association        4     73.2986    <.0001


Total Sample Size = 1362
```



```
▒alysis of Income Happiness data Midranks

The FREQ Procedure

Summary Statistics for Income by Happiness

  Cochran-Mantel-Haenszel Statistics (Based on Ridit Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
-----------------------------------------------------------------
    1        Nonzero Correlation        1     55.6873    <.0001

Total Sample Size = 1362
```