Solution to HW2

Problem 2.16

(a) The response variable is *lung cancer* and the explanatory variable is *smoking status*.

(b) This is a (matched) case-control study.

(c) No. Since this is a case-control study, the proportions who suffered lung cancer among smokers and non-smokers calculated from this study are NOT good estimates of the underlying true probabilities.

(d) Ignoring the matching, we can estimate the association using the odds-ratio:

$$\widehat{\theta} = \frac{688 \times 59}{21 \times 650} \approx 3.$$

Since the lung cancer is a rare disease, $\widehat{\psi} \approx \widehat{\theta} \approx 3$. Therefore, smokers are about 3 times as likely to have a lung cancer as non-smokers.

Problem 2.18

(a) $35.8 = (21 + 159 + 110) \times (21 + 53 + 94)/1362$.

(b) The $df = (3-1) \times (3-1) = 4$. The P-value$= P(\chi^2_4 \geq 73.4) = 4.3 \times 10^{-15}$.

(c) The standardized residual $z = -2.97$ in the cell with count 21 meaning that there are significantly less individuals who are not too happy but with income above average than predicted by the independence model.

The standardized residual $z = -5.907$ in the cell with count 83 meaning that there are significantly less individuals who are very happy but with income below average than predicted by the independence model.

(d) The standardized residual $z = 3.144$ in the cell with count 110 meaning that there are significantly more individuals who are very happy and have income above average than predicted by the independence model.

The standardized residual $z = 7.368$ in the cell with count 94 meaning that there are significantly more individuals who are not too happy and have income below average than predicted by the independence model.

(e) The SAS program and output for CMH1 are:

```
data prob2_18;
  input income happy count @@;
  datalines;
  1 1 21   1 2 159  1 3 110
  2 1 53   2 2 372  2 3 221
  3 1 94   3 2 249  3 3 83
;

title "Analysis of Income/Happy data";
proc freq data=prob2_18 order=data;
  weight count;
  tables income*happy / scores=rank cmh;
run;
```

Summary Statistics for income by happy

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

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	55.6873	<.0001
2	Row Mean Scores Differ	2	64.6145	<.0001
3	General Association	4	73.2986	<.0001

So $\chi^2_{CMH1} = 55.93, df = 1$ with P-value < 0.001.

(f) The SAS program and output for CMH2 are:

```
title "Analysis of Income/Happy data";
proc freq data=prob2_18 order=data;
  weight count;
  tables income*happy / cmh;
run;
```

Summary Statistics for income by happy

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

So $\chi^2_{CMH2} = 64.61, df = 2$ with P-value < 0.001.

Problem 2.21

(a) No, since an individual may appear in more than 1 cell.

(b) If we are interested in one factor, then it is straightforward to construct the required 2×2 table. For example, the table relating gender to factor A is

		Factor A	Factor $\bar{A}$	
Gender	Men	60	40	100
	Women	75	25	100

Problem 2.23 (sample report)

The Pearson chi-squared test for independence between *education* measured by the highest degree and *religious beliefs* gives $\chi^2 = 69.2$ with $df = 4$. Therefore there is strong evidence that *education* and *religious beliefs* are not independent (P-value $= 3.4 \times 10^{-14}$). The strong evidence against the independence model is also reflected in many large standardized residuals, especially in corner cells. The sign and magnitude of these standardized residuals may indicate the relation between *education* and *religious beliefs*. For example, the standardized residual 4.5 in cell (1,1) indicated that that there are many more individuals with less than high school education who are fundamentalist than predicted by the independence model; the standardized residual 6.5 in cell (3,3) indicated that that there are many more liberal individuals with Bachelor or graduate degree than predicted by the independence model; the change of the standardized residuals in column 1 from 4.5 to -6.8 indicates that with more education there are many less Fundamentalist individuals than predicted by the independence model; the change of the standardized residuals in column 3 from -1.9 to 6.3 indicates that with more education there are many more liberal individuals than predicted by the independence model. Since both *education* measured the highest degree and *religious beliefs* are ordinal categorical variables, the residual pattern indicates a "monotone" relationship between these two variables.

Problem 2.30

The P-value of Fisher's exact test for testing $H_0 : \theta = 1$ *v.s.* $H_a : \theta > 1$ is

$$\text{P-value} = P(n_{11} \geq 21 | \text{margins}, H_0) = \frac{\binom{23}{21}\binom{18}{15} + \binom{23}{22}\binom{18}{14} + \binom{23}{23}\binom{18}{13}}{\binom{41}{36}} = 0.3803,$$

indicating that there is no strong enough evidence against the null hypothesis at level 0.05 that surgery and radiation therapy are the same in controlling cancer given that surgery is no inferior to radiation therapy. However, this P-value is too conservative. That is, this test has a low power to reject H_0 even if H_a is true.

Problem 2.31

(a) Since under H_0 and given margins, the distribution of $n_1 1$ is

n11	18	19	20	21	22	23
p(n11)	0.045	0.213	0.362	0.275	0.094	0.011

The exact two-sided P-value for testing $H_0 : \theta = 1$ *v.s.* $H_a : \theta \neq 1$ can calculated

$$\text{P-value} = 0.045 + 0.213 + 0.275 + 0.094 + 0.011 = 0.638.$$

There is no enough evidence at level 0.05 against the null hypothesis that surgery and radiation therapy are the same in controlling cancer. However, this P-value is too conservative. That is, this test has a low power to reject H_0 even if H_a is true.

(b) One-sided mid P-value is:

$$\text{P-value} = 0.275/2 + 0.094 + 0.011 = 0.2425.$$

Using this one-sided mid P-value, we still do not reject $H_0 : \theta = 1$ at level 0.05. The advantage of using mid P-value is that the type-I error probability is closer to the nominal level (however, sometimes it can be greater).