Chapter 7

Categorical Data Analysis

Aim: Upon completing the chapter, the learner should be able to perform a stratified analysis in an epidemiological study, using cohort and case-control study designs.

7.1 Introduction

So far, we have discussed the Stata commands for estimating the conditional expectations of continuous variables. There are, however, numerous occasions in the public health field in which we are interested in exploring the association between a categorical outcome (e.g., disease status) and one or more predictor variables (e.g., exposure status, confounding variables, and effect modifiers variables) collected in epidemiologic studies. *Epidemiology* is "the study of the occurrence and distribution of health-related events in specified populations and the application of this knowledge to control relevant health problems" (Porta, 2008; Rothman, 2002). Epidemiological studies are commonly categorized as descriptive or analytical studies. These studies are defined immediately below:

- *Descriptive epidemiology* focuses on describing the occurrence (incidence, prevalence, and mortality) and distribution of disease (or other health event) patterns by characteristics relating to person (who is affected by the health event?), time (when does the health event occur?), and place (where does the health event occur?). Descriptive studies often use routine data (i.e., vital statistics, surveillance systems, registries, or population surveys) collected in a population to characterize the patterns of disease occurrence. The data generated from descriptive studies can be used for healthcare planning and hypothesis generation. Types of descriptive studies include case series, cross-sectional, and ecological studies.

- *Analytical epidemiology* is concerned with assessing the associations between exposures and diseases (or other health outcomes), which associations may provide further insights into the causes of a disease and lead to prevention strategies. Types of analytical studies include case-control studies, cohort studies, and clinical trials.

In the next sections, we will show the application of the Mantel–Haenszel method for the analysis of data derived from cohort and case-control studies (Jewell, 2004; Rothman, 2002). This method is based on the stratification of potential confounding variables to estimate a weighted average of the magnitude of the exposure–disease association. Confounding factors are variables that are related to both the exposure and the outcome but do not lie in the causal pathway between them (Rothman, 2002; Woodward, 2004). As we shall see in the next chapters, regression models are efficient techniques that can be employed to assess the exposure–disease association while controlling for the *confounding variables*.

7.2 Cohort Study

Cohort studies are designed with at least two groups of subjects usually called *exposed* and *nonexposed* (in terms of a particular factor, in either case) groups. Once these groups are identified, they are followed up for a specific period of time to determine whether any of their members develop the disease of interest, while controlling for potential confounding variables. The magnitude of the association between *exposure* and *disease* is determined by the *relative risk*, which is defined in the following way:

$$\mathrm{RR} = \frac{I_{\text{exposure}}}{I_{\text{nonexposure}}}$$

where I_j indicates the incidence of the disease in the jth group. When stratified analysis is performed, the RR is assessed under different strata (to be combined into one single RR) or reported in each stratum. In the *Mantel–Haenszel* method, the combined RR is computed using the weighted mean of the RRs, as follows:

$$\mathrm{RR}_{\mathrm{M\text{–}H}} = \sum \frac{w_k \mathrm{RR}_k}{w_k}$$

where w_k is the weighted factor in the kth stratum, which is itself determined with the product of the total number of cases who are unexposed and the proportion of exposure in this stratum.

For example, let us say that we want to evaluate the association between alcohol intake (exposure) and a diagnosis of myocardial infarction (MI) over a period of 5 years, controlling for the effect of cigarette smoking (potential confounding variable). To analyze this type of study in Stata, we can use the following data:

	Smoker (0)		
Alcohol	MI + (1)	MI – (0)	*Total*
Present (1)	8	16	24
Absence (0)	22	44	66
Total	30	60	90
	Nonsmoker (1)		
Alcohol	MI + (1)	MI – (0)	*Total*
Present (1)	63	36	99
Absence (0)	7	4	11
Total	70	40	110

Between the parentheses are the codes for each category (1 indicates presence and 0 indicates absence).

To program these data, the database can be entered in Stata as follows:

```
     +-----------------------------------+
     | smoker   alcohol   mi   subjects |
     |-----------------------------------|
  1. |      0         1    1          8 |
  2. |      0         1    0         16 |
  3. |      0         0    1         22 |
  4. |      0         0    0         44 |
  5. |      1         1    1         63 |
  6. |      1         1    0         36 |
  7. |      1         0    1          7 |
  8. |      1         0    0          4 |
     +-----------------------------------+
```

The command to perform a stratified analysis using the Mantel–Haenszel method is *cs*, as is illustrated in the following:

```
cs mi alcohol [fw=subjects], by (smoker)
```

Output

```
        smoker |      RR       [95% Conf. Interval]    M-H Weight
---------------+--------------------------------------------------
             0 |         1    0.5164877   1.936154      5.866667
             1 |         1    0.6244517   1.601405           6.3
---------------+--------------------------------------------------
         Crude |   1.53266      1.10769   2.120674
  M-H combined |         1    0.6695272   1.493591
---------------------------------------------------------------
Test of homogeneity (M-H)   chi2(1) = 0.000   Pr>chi2 = 1.0000
```

Note: When the database collapses and contains a variable that tells the frequency of each observation, the *fw* option is used. This option specifies the variable that contains the number of times the observation was actually observed.

The output reports the point estimation of the relative risk (RR) for each stratum, as well as the crude RR and the weighted RR (RR_{M-H}) with their respective 95% confidence intervals. In addition, the weighted factor in each stratum is reported (M–H weight), as is the significance test (test of homogeneity [$H_0: RR_1 = RR_2 = \cdots = RR_k$]), to assess whether the RRs in all strata are equal.

The results indicate that there is a nonsignificant difference in the RRs, per stratum (P-value > .10); therefore, it is recommended that the RR_{M-H} be used. When we compare the point estimates of the crude $\widehat{RR} = 1.53$ and the $\widehat{RR}_{M-H} = 1$, we are able to conclude that the data show a strong confounding effect, as the crude RR is overestimating the magnitude of the association between MI and alcohol intake. Finally, the estimated magnitude of the association of interest, controlling for smoking, is 1 (95% CI: 0.67, 1.49); this, however, is not statistically significant (P-value > .05).

7.3 Case-Control Study

Case-control studies are designed initially with at least two groups of subjects, called *cases* (diseased) and *controls* (nondiseased). Once these groups are identified as having been exposed (or not) to a specific factor, they are then classified as being either exposed or nonexposed groups. The magnitude of the association between exposure and disease is determined by the *odds ratio* (OR), which is calculated with the following expression:

$$\text{OR} = \frac{\text{Odds}_{\text{exposure}}}{\text{Odds}_{\text{nonexposure}}}$$

where Odds_j indicates the expected number of cases per control in the jth group (exposed or nonexposed) and can be defined as follows:

$$\text{Odds} = \frac{p}{1-p}$$

where p is the probability of having a diagnosis of the disease of interest under the study design.

In the Mantel–Haenszel method, the combined OR is computed using the weighted mean of the ORs, as follows:

$$\text{OR}_{M-H} = \sum \frac{w_k \text{OR}_k}{w_k}$$

where w_k is the weighted factor in the kth stratum, which is determined with the product of the number of cases who are unexposed and the number of controls who are exposed divided by the number of subjects in this stratum.

For example, let us assume that the user wants to evaluate the association between HPV (human papilloma virus) infection status and oropharyngeal cancer (OC), stratified by smoking (smokers vs. nonsmokers), using a case-control design with the following data:

Smoker (0)			
HPV	OC + (1)	OC – (0)	*Total*
Present (1)	**75**	**20**	**95**
Absent (0)	**5**	**80**	**85**
Total	**80**	**100**	**180**
Nonsmoker (1)			
HPV	OC + (1)	OC – (0)	*Total*
Present (1)	**5**	**18**	**23**
Absent (0)	**10**	**72**	**82**
Total	**15**	**90**	**105**

To perform the stratified analysis of these data, the database in Stata is prepared as is seen here:

```
. list
     +-------------------------------+
     | smoker   hpv   oc   subjects |
     |-------------------------------|
  1. |      0     1    1         75 |
  2. |      0     1    0         20 |
  3. |      0     0    1          5 |
  4. |      0     0    0         80 |
  5. |      1     1    1          5 |
     |-------------------------------|
  6. |      1     1    0         18 |
  7. |      1     0    1         10 |
  8. |      1     0    0         72 |
     +-------------------------------+
```

Using the Mantel–Haenszel method, the following command can be employed to perform the stratified analysis in a case-control study:

```
cc oc hpv [fw=subjects], by(smoker)
```

Output

```
        smoker |       OR       [95% Conf. Interval]  M-H Weight
---------------+-------------------------------------------------
             0 |         60       20.21104   207.9978    .5555556 (exact)
             1 |          2       .4721323   7.399327    1.714286 (exact)
---------------+-------------------------------------------------
         Crude |   21.33333       10.63312   43.91911             (exact)
  M-H combined |    16.1958       8.529819   30.75142
-------------------------------------------------------------------
Test of homogeneity (M-H)   chi2(1) = 18.06   Pr>chi2 = 0.0000

                  Test that combined OR = 1:
                              Mantel–Haenszel chi2(1) = 89.05
                                                Pr>chi2 = 0.0000
```

The output reports the point estimation of the OR for each stratum as well as the crude OR and the weighted OR (M–H combined) with 95% confidence intervals, respectively. In addition, the weighted factor in each stratum is reported (M–H weight) as well as two significance tests. The purposes of these tests are as follows:

1. To assess whether the ORs in all strata are equal: *test of homogeneity* ($H_0 : \mathrm{OR}_1 = \mathrm{OR}_2 = \cdots = \mathrm{OR}_k$).
2. To assess whether the weighted OR is equal to 1: *test of combined* OR = 1 ($H_0 : \mathrm{OR}_{\mathrm{M-H}} = 1$).

The results indicate that there is a significant difference in the ORs of each stratum (P-value < .05); therefore, it is recommended that the OR be analyzed per stratum. When we compare the point estimates of the OR_0 (60) in nonsmokers and those of the OR_1 (2) in smokers, we can see that the smoking habit modifies the magnitude of the association between HPV and OC. Finally, the estimated magnitude of the association of interest among smokers is 60 (95% CI: 20.2, 207.9), which is statistically significant (P-value < .05).

7.4 Sample Size and Statistical Power

To determine the minimum sample size for assessing the null hypothesis, H_0: OR = 1, with enough statistical power (i.e., $1 - \beta = 0.8$) and a minimum level of significance (i.e., $\alpha = 0.05$), the classical formula for comparing two proportions is recommended (Rosner, 2010). In Stata, this can be performed in the option that features the chi-squared test comparing two independent proportions with the *Power and sample size analysis* window in the *Statistics* menu (Figure 7.1).

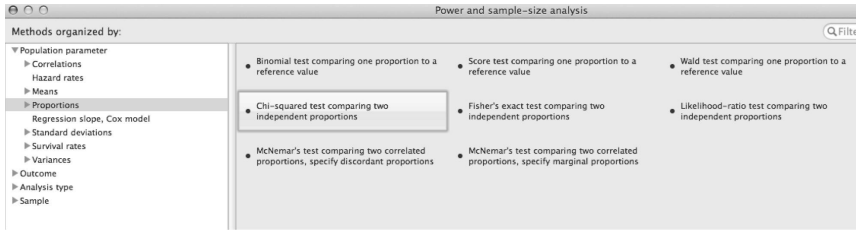

Figure 7.1 Sample size for comparing two independent proportions.

For example (using the data of nonsmokers), to assess the magnitude of the association between HPV status and OC with ORs = 2, 2.5, and 3, assuming that the prevalence estimates of OC in HPV-negatives are 0.10, 0.15, and 0.2, the table of sample size should be filled in for a one-sided test and equal allocation, as illustrated in Figure 7.2.

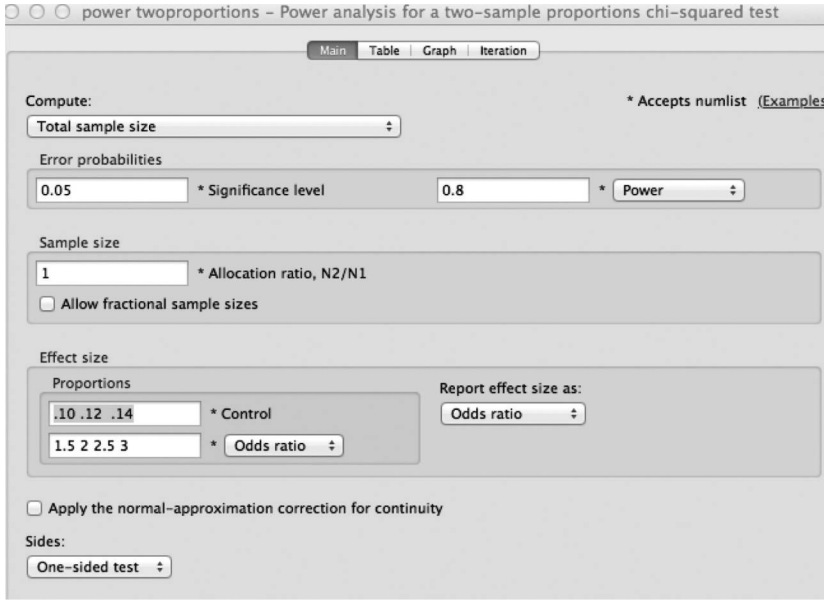

Figure 7.2 Sample size specifications for comparing two independent proportions under different conditions.

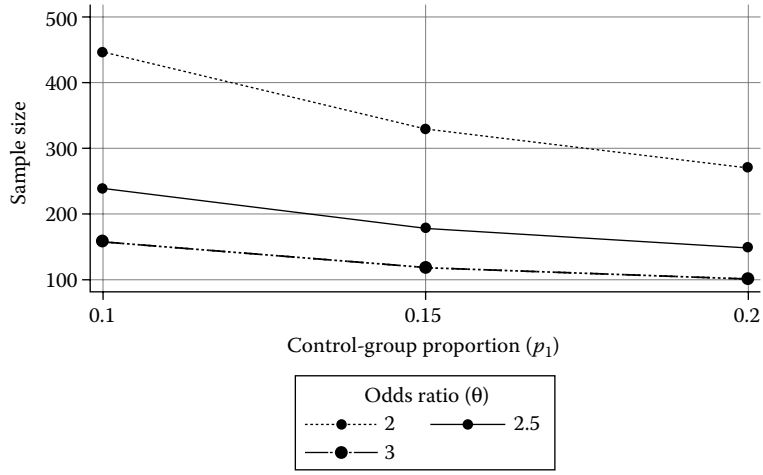

Figure 7.3 Alternatives of sample size for comparing two independent proportions under different conditions. Parameters: $\alpha = 0.05$, $1 - \beta = 0.8$.

Once the window for the previous requested sample size is submitted with the graph option, a plot is displayed (Figure 7.3). The results show that the lower the OR, the total sample size increases; however, when the proportion of OC in HPV-negative groups is incremented, the differences in sample size are reduced. To determine the sample size for the overall OR while controlling for potential confounders, an adjustment has to be made, as explained in Chapter 8 (Hosmer and Lemeshow, 2000).