

Elementary Epidemiological Data Analysis Using STATA

By

Robert A. Yaffee, Ph.D.
Statistics, Social Science, and Mapping Group
Academic Computing Services
Information Technology Services
New York University
December 2001

Keywords and Key Phrases: attributable risk, case, case-control study, cohort study, exposure, exposure time, incidence rate, risk, risk difference, risk factor, relative risk, odds ratio.

Introduction

STATA permits the analyst to perform elementary epidemiological analysis. To study the nature of disease (morbidity) or fatality from a disease (mortality), epidemiologists conduct **cohort or case-control studies**. With simple, one-line commands, such a researcher can generate the preliminary tables and associate computations used for such analysis. These tables can be either 2 by 2, 2 by K tables, or stratified tables. When the analysis has dichotomous outcomes, STATA can output the 2 by 2 tables. When the analysis has ordinal outcomes, STATA can produce the 2 by K tables for this analysis. When the analysis requires comparison of tables, STATA can produce the stratified tables. In this paper, an introduction into the preliminary tabular analysis employed by epidemiologists is presented. To begin, the research design of the incidence rate, cohort and case-control study are presented. After reviewing the basic design of these studies, this article explains the data formats, the command syntax, along with the statistical analysis for the epidemiological study.

Types of Research Designs

Epidemiological observational research consists largely of cohort and case-control studies. Cohort studies can be either historical or prospective (Breslow and Day, 1980). In the historical cohort study, the medical history of a cohort of subjects is reconstructed with a view toward defining a cohort of those who were exposed to a risk factor. The rate of those subjects who exhibit the disease or death among those exposed is compared with the rate of disease or mortality among an unexposed control group. Sometimes, the population rate is used as a basis of comparison.

The historical cohort study has particular requirements. A group exposed to a risk factor must be identified. The historical exposure data must be readily available. There must be a clear time-frame to the study. There must be a clear time of onset of exposure. Where this may differ from person to person, the exposure time may be measured in person-time. The sample size and level of exposure needed to yield significant results is always an issue with these studies.

There are problems with the historical cohort studies. The reconstruction of the medical history can be difficult when records are inaccessible or incomplete. If patient recall is used, then recall bias can afflict the study. If there is insufficient data on the potentially confounding variables, this may undermine the validity of the results. This lack of other alternative exposures is a common problem with retrospective cohort studies. Information about the health of the groups prior to the detection of the disease may be difficult to obtain. It is often difficult or impossible to assess selection bias in this sort of study. It is even more difficult to eliminate it.

In a prospective cohort study, the subjects who pass the preliminary screening are observed over a period of time. The potential subjects must be screened to determine who can be included in the study. The subjects eligible to participate are screened to eliminate those afflicted with problems whose symptoms would confound the clear identification of a case. Those who pass the inclusion-exclusion criteria are identified as the cohort to be followed. One group in this cohort is exposed to a risk factor whereas others are not. During this study, the incidence rate of those exposed is compared with those unexposed. The incidence rate is the rate at which persons without the disease--namely, noncases-- are transformed into persons with the disease--namely, cases.

There are specific requirements of this kind of study. It must be determined who can be admitted to the study and who among those who apply must be excluded. How the dates of entry and exit are determined must be decided. The procedures of observation must be clearly defined. There must be a clear notion of whether and what kind of exposure was sustained. In other words, exposure status must be easily and clearly determined. The period of observation must be clearly defined and measured. By clearly identifying the nature of the disease, it must also be clear how to detect its onset and to identify the cases. The follow-up mechanisms must be agreed upon. If a professional or occupational group is being followed, then definitions of membership must be clearly established. Similarly, definitions of an exit via retirement or loss of employment must be established. There must be clearly agreed upon procedures when migration, accidental injury or demise brings about departure from the group being followed. The control group must be clearly defined. The population must be clearly determined. In some cases, there can be several comparisons or control groups.

There are disadvantages of this kind of study. The careful following of subjects over an extended time-span is required to determine the extended commitment of dedicated professions, funds, and time. Tracking patients over time can require much effort and patient attrition for a variety of reasons over an extended span of time can be a problem. The advantages of this kind of study are that recall bias can be eliminated. Sometimes selection bias can be eliminated. Potentially confounding effects can be eliminated. Biological data affecting the health of the subjects prior to the detection of the disease may be available.

In 1951, the British doctors began a prospective cohort study of the association between smoking and the incidence of lung cancer. In 1954, Doll and Hill reported the comparison of the lung cancer rate of smokers to that of nonsmokers, for example. They needed a complete smoking history of their subjects. The Life Span Study of the atomic bomb survivors is a study where the proportional mortality of the survivors was compared with that of those relatively unexposed to the radioactive fallout. In this way the incidence rate of the exposed is directly observed and then compared to that of those in some control group (Breslow and Day, 1987).

The **incidence rate** is the focus of many cohort studies. The **incidence rate** is the rate at which persons without the disease (known as noncases) are transformed into persons with the disease (designated as cases). Two groups of people are followed over a period of time. One group is exposed to a risk factor and the other group is not. The exposure of the person over time, called person time, is measured uniformly in hours, days, months or years from the time of the beginning of the exposure to the endpoint of the exposure under study. The person time per group is measured by the sum of the exposure times for all members in the group. In incidence rate studies, the proportion of cases per cumulative person-time, for those in the exposed group, is compared to that in the unexposed group.

Whether there is a significant difference between the incidence rate of the exposed group, versus that of the unexposed group, can be assessed. STATA can compute the incidence rates of the two groups and their difference, known as the risk difference. The calculation of the confidence intervals allows the analyst to determine whether this difference is due to chance. STATA also computes the relative risk, by dividing the incidence rate of the exposed group by that of the unexposed group. Moreover, STATA calculates the proportion of the cases that are attributable to exposure as well as the net proportion of cases in the whole population attributable to exposure. In these ways, STATA permits the epidemiologist to analyze the incidence rate of the disease. STATA does not limit itself to incidence rate cohort studies.

In the retrospective **case-control study**, the basis for selection of the two groups is whether the subject had the disease or not. This study begins with the determination of who has the disease and who does not. The persons with the disease are called cases and the persons

without the disease are called controls. Rigorous inclusion and exclusion criteria are established. Subjects with potentially confounding etiological factors or symptoms are excluded. The exposure histories of the subjects over a well-defined period of time are obtained and reconstructed. The association of the exposure and the disease is analyzed. The case-control study typically involves a smaller number of more accessible subjects than the cohort study, but may be vulnerable to recall bias or partial historical reconstruction. Sometimes previous errors in measurement may bias the assessment of the exposure or proportionate mortality. This problem has to be guarded against. The question of whether occupational, hospital, or population controls are used becomes an issue and which sources of data are adequate under these circumstances. Although these problems may produce selection bias, this kind of study is cheaper and speedier to complete than the cohort study. For these reasons, the case-control study is the more popular of the two types of observational studies (Breslow and Day, 1980).

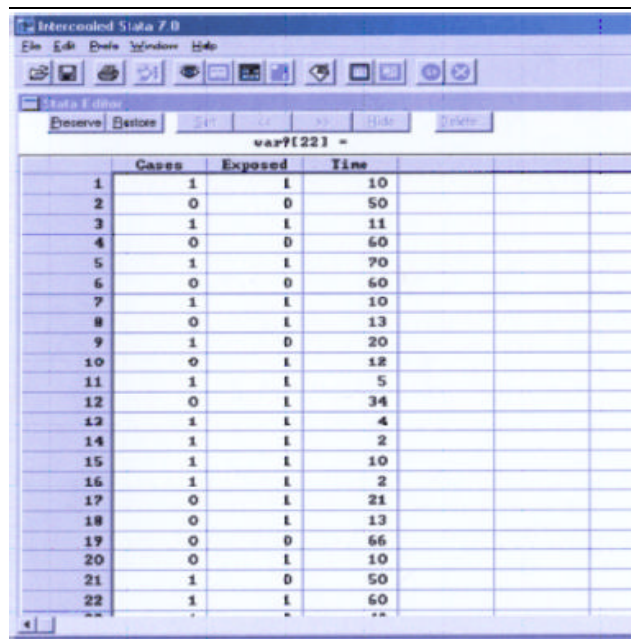
Hypothesis Construction

The research hypothesis stipulates that under specified conditions, the risk factor is associated in some manner with the onset of disease, syndrome, or death. A hypothesis can be constructed relating the incidence of a disease to exposure to a particular chemical or mineral found in the environment. The disease and exposure are operationalized to permit unambiguous measurement of them. A statistical test of the hypothesized relationship between these indicators is then applied to confirm or disconfirm the hypothesis.

In the cohort study, the independent variable is the risk and the dependent variable is the disease outcome. A cohort, a group of persons who share a common experience within a defined period of time (Mausner and Kramer, 1985), is selected according to whether it can be diagnosed according to the disease during the time of observation. The cohort study involves the observation of at least two groups of subjects, which differ in the amount of exposure to a risk factor, over time. They are used when the disease is not particularly rare or where there are several occupations or substances that are associated with the disease (Mausner and Kramer, 1985). In the incidence rate study, the hypothesis deals with the association between exposure and number of new cases of the disease within the time period. More specifically, the hypothesis focuses on the incident rate difference (or an incidence rate ratio) between the exposed and unexposed groups. It may stipulate that this difference or ratio is statistically significant, and that it will be a fairly large small, moderate, or large one. In the cohort, study, the hypothesis may stipulate that there will be a statistically significantly greater number of cases among those exposed to the risk factor than among those who are not. Alternatively, it may stipulate that there will be a statistically insignificant difference between cases and noncases among those who exposed to the risk factor.

Data File Format

There are two basic data entry formats for 2x2 tabular epidemiological analyses with STATA. The first data format consists of a regular rectangular data file. In this kind of file, each horizontal line represents an individual case and each column represents a separate variable. There are as many lines as there are cases and there are as many columns as there are variables. In an incidence rate study, the three variables needed for each case are whether the subject turned out to be a case or noncase and whether the subject was exposed or unexposed and how much time was he exposed to the risk factor. This format appears in Figure 1.



The screenshot shows the STATA Data Editor window with a dataset named 'was91221'. The data is organized into three columns: 'Cases', 'Exposed', and 'Time'. The rows represent individual subjects, numbered 1 through 22. The 'Cases' column contains binary values (0 or 1), the 'Exposed' column contains binary values (0 or 1), and the 'Time' column contains numerical values representing exposure duration.

	Cases	Exposed	Time
1	1	1	10
2	0	0	50
3	1	1	11
4	0	0	60
5	1	1	70
6	0	0	60
7	1	1	10
8	0	1	13
9	1	0	20
10	0	1	12
11	1	1	5
12	0	1	34
13	1	1	4
14	1	1	2
15	1	1	10
16	1	1	2
17	0	1	21
18	0	1	13
19	0	0	66
20	0	1	10
21	1	0	50
22	1	1	60

Figure 1: Rectangular Data Format with Rows as subject and Columns as variables

With this kind of data file input format, the command the analyst would enter on the STATA command line to produce a 2 by 2 cross tabular analysis of incidence rates is

ir Cases Exposed Time

The output that this command generates can be seen in Figure 2.

Stata Results			
. lr Cases Exposed Time			
	Exposed to risk factor		Total
	Exposed	Unexposed	
Disease count	15	5	20
Person Years	476	621	1097
Incidence Rate	.0815126	.0080515	.0182915
	Point estimate		[95% Conf. Interval]
Ino. rate diff.	.0294611	.006022	.0409002
Ino. rate ratio	3.913866	1.352139	13.76406 (exact)
Attr. frac. ex.	.7444981	.260428	.927347 (exact)
Attr. frac. pop	.5589736		
	(midp) Pr(k)=15 =		0.0026 (exact)
	(midp) 2*Pr(k)=15 =		0.0052 (exact)

Figure 2

If the data are available in tabular matrix format, the following STATA command written in the STATA command editor will generate the proper output. Suppose that the counts in each cell of the matrix, shown in Table 1 below, are labeled, A through D, from left to right along the top row and then along the bottom row. STATA is case-sensitive so, the command must be written in noncapital letters. Then the general form of the command to analyze those results is:

lri A B C D

Table 1 Entering Data in Tabular Format			
	Exposed	Unexposed	Marginal Totals
Disease Count	A	B	P
Person Years	C	D	T
Totals	$N = A + C$	$M = B + D$	

In this particular case, the command would be:

iri 15 5 476 621

The output that this command generates is shown in Figure 3

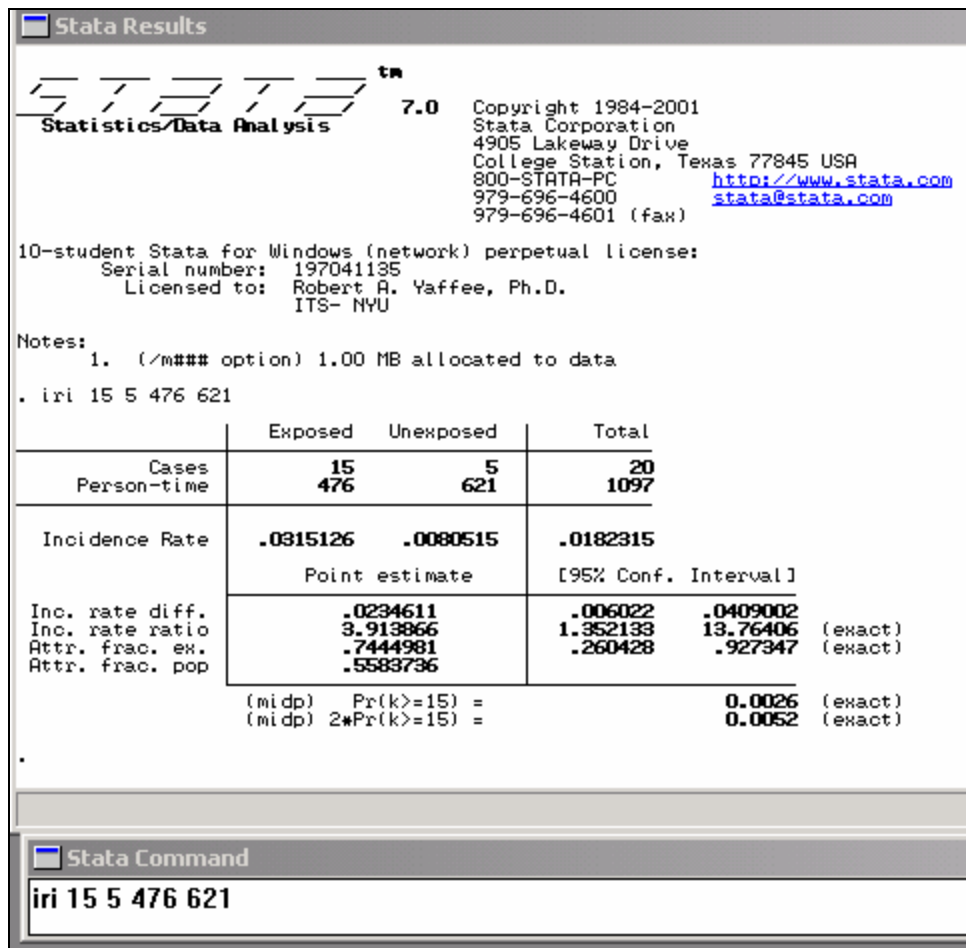


Figure 3 Incidence Rate Analysis with Tabular Data Input

Incidence Rate Analysis

The incidence rate analysis output is the same as that shown in Figure 2. In a hypothetical example, we might decide to test of the effects of arsenic in the drinking water. Suppose there is a question among residents of a given area about the health hazard facing the public from arsenic

pollution of the water supply. From studies of Montana smelter workers, higher concentrations of airborne arsenic have been found to be associated with respiratory cancer (Lee and Fraumeni, 1969). Concerned persons contend that from ingestion of arsenic, gastrointestinal irritations, difficulty with swallowing, low blood pressure, and convulsions may result. Over the longer term, there is an increased danger of skin, liver, lung, bladder, and kidney cancer. Meanwhile, suppose other more cavalier persons contend that there is plenty of naturally occurring arsenic in seafood and ground water and that there is not enough of it in the well water to constitute a health hazard that is worth the cost of filtering their water.

To test the hypothesis that extended exposure to arsenic pollution of the drinking water is a carcinogenic health hazard, the people who had been living in one locale with heightened levels of arsenic in their drinking water for 10 years were deemed to have been exposed. Those living in a distant community with reduced levels of arsenic in the drinking water for the same period of time were deemed to be the control group. The researchers retrospectively observe medical history of these groups over the defined period of time. The time of exposure for each person is recorded. At the end of the study, a search for a case of cancer of the skin, liver, lung, bladder, or kidney is undertaken. If any of these cancers were diagnosed, the person was deemed to be a case. Otherwise, the person was deemed to be a noncase. An incidence rate analysis is conducted.

Therefore, we gather the data and run the analysis. The command by which the arsenic concentration is tested comes from a cross-tabulation of the cases and person-time by those exposed.

Statistical Interpretation

With an incidence rate analysis, those with the disease are divided into the exposed and unexposed groups. Their counts are inserted into cells a and b. Their person-time of those exposed is summed and that of those unexposed is also summed. These totals are inserted into cells c and d. The ratio of exposed cases to exposed person-time provides the first **incidence rate** (IR_{eg}). The ratio of the unexposed cases to the unexposed person-time provides the second incidence rate (IR_{ug}). If one takes the second rate and subtracts it from the first, one obtains the point estimate of the excess risk or **incidence rate difference** (IR_d).

$$IR_d = IR_{eg} - IR_{ug} \quad (1)$$

If one divides the second rate into the first incidence rate, one obtains the point estimate of the **incidence rate ratio** (**IRR**). This ratio is also called the **relative risk** (**RR**):

$$RR = \frac{IR_{eg}}{IR_{ug}} \quad (2)$$

The magnitude of the relative risk indicates the strength of the relationship between exposure and incidence. A relative risk of unity indicates that there is no difference between the incidence risk from the exposure group than that from the control group.

A relative risk greater than unity indicates that exposure is related to higher incidence: The higher the relative risk, the more the association between the exposure and the incidence. Correspondingly, a relative risk less than unity indicates that exposure is associated with lower levels of incidence. Relative risks less than unity suggest that exposure is associated with a reduced incidence of the disease than would be found in the control group. Relative risks less than unity are often associated with therapeutic drugs used to thwart morbidity or mortality.

Not only are the point estimates produced, the exact interval estimates are output. The **confidence limits** for the incidence rate difference and ratio are computed. Their statistical significance depends on whether these confidence interval brackets zero. If, on the one hand, the confidence limits bracket zero, then the statistical difference or ratio is insignificant. If, on the other hand, these limits do not bracket zero, the incidence rate difference or incidence rate ratio is statistically significant.

The confidence limits of the incidence rate are easily formed. First one computes

$$v = \frac{A - CP / T}{\sqrt{PCD / T^2}} \quad (3)$$

The confidence interval for the incidence rate difference is formed by

$$IR_d \pm \frac{z}{v} \quad (4)$$

and the confidence interval for the incidence rate ratio is formed by

$$RR^{1 \pm \frac{z}{v}}$$

where n is defined in Eq.3 and

$z = z$ score for confidence interval.

(5)

If a representative sample was conducted, it is possible to ascertain the unbiased estimate of the distribution of exposure levels in the population. From that, it is possible to ascertain the extent

to which the distribution of cases in the population is due to exposure. Given the proper sampling, STATA can compute the **attributable risk for exposed persons**. This fraction of the incidence rate attributable to exposure, as well as its confidence limits, is output. The proportion of cases in the difference between exposed minus the unexposed divided by the proportion of cases among the unexposed is called the fraction of the incidence rate attributable to exposure. This fraction is placed below the incidence rate ratio to the left of the tabular output.

$$\begin{aligned} & \textit{Fraction Attributable to Exposure} \\ & = \frac{IR_{eg} - IR_{ug}}{IR_{eg}} \end{aligned} \quad (6)$$

Where IR_{eg} = incidence rate of exposed group
 IR_{ug} = incidence rate of unexposed group

In addition to attributable risk for exposed persons, the **population attributable risk (PAR)** is output by STATA. The population attributable risk is the proportion of cases occurring in the total population which can be explained by the risk factor.

To obtain proportion of cases in the total population explained by the risk factor, **the population attributable risk**, the net proportion of all cases in the population attributable to exposure, is computed. To obtain the numerator for the ratio, the proportion of persons in the population exposed to the risk factor is multiplied by the risk difference. To obtain the denominator for the ratio, the proportion of persons in the population exposed to the risk factor is multiplied by the incidence rate of the exposed group and add that to the product of the proportion of persons in the population exposed to the risk factor by the incidence rate of the unexposed. This ratio is the population attributable risk. The formula for it can be found in Breslow and Day, 1980.

$$PAR = \frac{p(IR_{eg} - IR_{ug})}{p(IR_{eg}) + (1-p)IR_{ug}} \quad (7)$$

where

AR = Population Attributable Risk

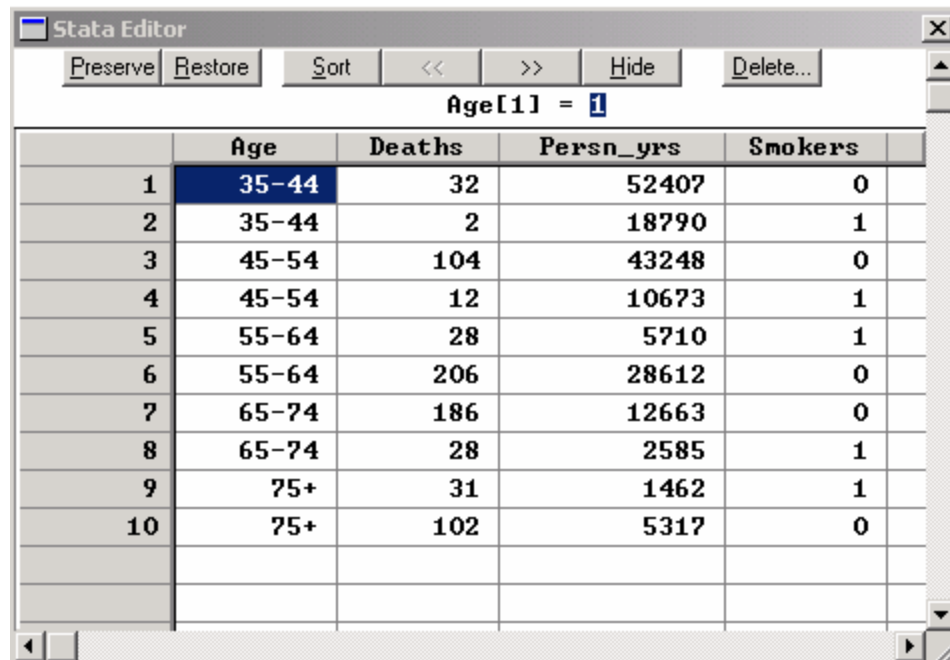
p = the proportion of persons in the population exposed to the risk factor

IR_{eg} = the incident rate for the exposed subgroup

IR_{ug} = the incidence rate for the unexposed subgroup

Stratified Incidence rate Analysis

Sometimes it behooves the researchers to examine whether a third variable is significantly related to the outcome. This variable may be an intervening or antecedent variable and it may be contended that this other variable is critical to the outcome. The Table can be stratified by this third variable. When the data are entered in such a way that they are stratified by an extra variable, then a stratified incidence rate analysis can be conducted. The data are from the smoking research from the British Doctor's study, reported by Doll and Hill (1966) and found in the *STATA7* Reference Guide (A-G), (see Figure 4 below) so that age, smoking, deaths, and person years of exposure are the variables, shown below.



Stata Editor

Age[1] = 1

	Age	Deaths	Persn_yrs	Smokers
1	35-44	32	52407	0
2	35-44	2	18790	1
3	45-54	104	43248	0
4	45-54	12	10673	1
5	55-64	28	5710	1
6	55-64	206	28612	0
7	65-74	186	12663	0
8	65-74	28	2585	1
9	75+	31	1462	1
10	75+	102	5317	0

Figure 4 Data Format for Stratified Incidence Rate Analysis, Data are from British Doctors Study Doll and Hill (1966) cited in STATA 7 Reference Guide (A-G), p.455.

To command STATA to perform the stratified incidence rate analysis, the "ir" command is entered into the Syntax command window:

ir Deaths, Smokers Persn_yrs, by (Age)

The output from this command is shown in Figure 5 below.

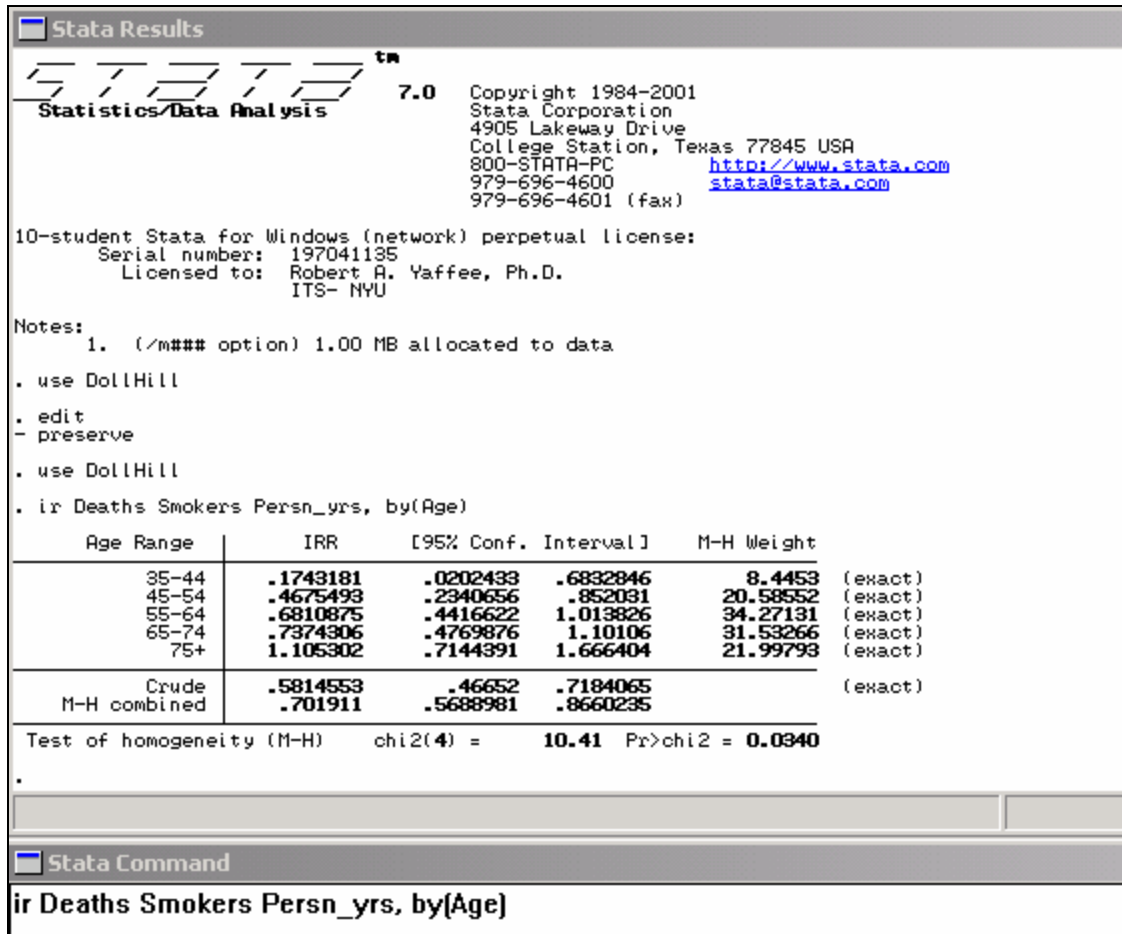


Figure 5 Stratified Incidence Rate Analysis of British Doctors Study

The application of this command produces the **incidence rate ratio** for each age. Along with each ratio, a set of confidence limits is output. For each age stratum in Figure 5, there is an exact M-H weight W_I computed.

$$W_i = B_i \text{Meg}_i / T_i \quad (8)$$

Where W_i = Mantel -Haenszel weight for stratum I
 B_i = Unexposed count for stratum I
 Meg_i = Total of Exposed groups for stratum I
 T_i = Total of strata over the strata

The combined incidence rate ratio (CIRR) is formulated as

$$CIRR = \frac{\sum_i a_i M_{ug_i} / T_i}{\sum_i W_i} \quad (9)$$

Where

- a_i = Count of exposed group for strata i
- M_{ug_i} = Total of unexposed groups for stratum i
- T_i = Total over strata i
- W_i = M-H weight for stratum i

The statistical test is performed with the standard error of the combined M-H incidence rate.

$$SE_{imh} = \sqrt{\frac{\sum_i P_{iMiNi} / T^2}{\left(\sum_i a_i M_i / T_i\right) \left(\sum_i b_i N_i / T\right)}} \quad (10)$$

The confidence intervals are constructed with this standard error and a combined Mantel-Haenszel test of the homogeneity of the strata is conducted. Because the $\text{Pr} > \chi^2 < .05$, the strata are inferred to be statistically heterogeneous. In other words, age does make a statistical difference in the distribution of the incidence rate ratios. Among these subjects in Figure 5, the **relative risk (incidence rate ratio { IRR })** seems to increase significantly with age.

The researcher may obtain either internal or external standardization of his results. If he wishes to obtain the **standardized mortality ratio (SMR)**, he can standardize his results by weighting each age group by the population of the exposed group. The SMR is defined as the total number of deaths divided by the expected number of deaths (Selvin, 1996). To do so, he stratifies the table with the **'by(age)'** option and obtains the requested standardization with the **istandard** option added to the command. If the researcher wishes to externally standardize with weights proportional to the population of the unexposed group, he can use instead the **estandard** with his incidence rate command.

ir Deaths, Smokers Persn_yrs, by (Age) estandard

Cohort Study Analysis

In a cohort study, two groups are followed over time and at the end of the study are compared for respective proportions that develop a particular disease-for example, respiratory cancer. The researchers compare the percentage that develops lung cancer over a five-year period of the experimental (the higher arsenic level) with that of the control (reduced arsenic level) group. These percentages are the incidence rates of the two groups. The incidence rates are

compared to assess the association of the risk factor with the incidence of the disease. The relative risk is defined as the incidence rate among the exposed divided by the incidence rate among the unexposed group.

The cross-tabulation of the model places the dependent variable, the amount of exposure in the columns and the independent variable, the disease outcome in the rows. The risks in the two groups are measured by the percent of the cases in the respective groups. After the risk of exposed group and the risk of the unexposed group are computed, the difference between these incidence rates is calculated. The risk ratio of the exposed to the unexposed is then computed along with its 95 percent confidence intervals.

Suppose that cells a, b, c, and d are counts of the number of subjects within the designated cell of the above table. The first row contains the 75 cases with the disease under study. Fifty-five of who were exposed to the risk factor. The second row contains 125 noncases (persons without the disease), 45 of whom were exposed to the risk factor. Whenever the table contains less than 1000 cases, it is recommended that the user should specify the **exact** option, which invokes the Fisher's Exact test of significance. The command typed into the command line areas is:

csi A B C D, exact

The input of this command yields the tabular output on the following page.

(subjects who have a disease) are selected and a group of controls (subjects who do not have a disease) are selected. The extent to which each group was exposed to a risk factor is investigated. These proportions of those with the disease are the incidence rates of the groups. The incidence rate of the unexposed group is subtracted from that of the exposed group to obtain the risk difference. The relative risk is formed by dividing the incidence rate of the unexposed group into that of the exposed group. The purpose is to ascertain whether the incidence rate of those exposed to the risk factor is statistically significantly greater than the percentage of those not exposed. A classification table with case-control information may be constructed with

cci 175 207 2825 6793, level(99)

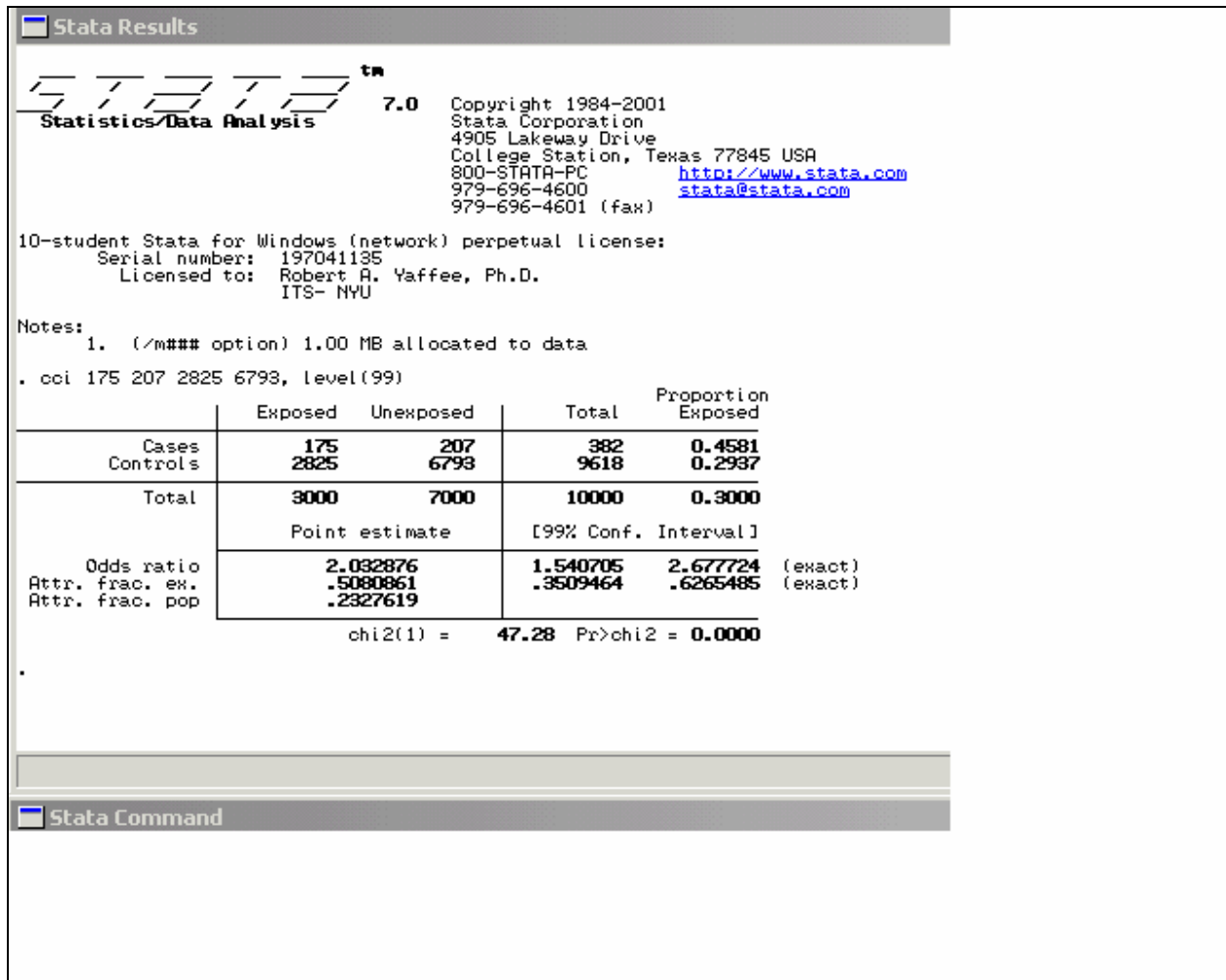


Figure 7 Output of Case-Control Analysis with Tabular Data Input

For the case group and the control group, the proportion exposed is listed next to the total in each group. The odds ratio (OR), also called the relative risk (RR), is computed as

$$\begin{aligned}
 \text{Odds ratio} &= \frac{\text{prob}(\text{case}) / (1 - \text{prob}(\text{case}))}{\text{prob}(\text{control}) / (1 - \text{prob}(\text{control}))} \\
 &= \frac{A / B}{C / D}.
 \end{aligned}
 \tag{11}$$

The confidence intervals are given as

$$OR^{1 \pm \frac{z}{v}}
 \tag{12}$$

where z = z score for confidence interval
 v = defined in Equation 3.

For those studies with representative samples, attributable fraction due to exposure and the attributed fraction of the population are also given. The χ^2 test with 1 degree of freedom reveals indication of a statistically significant relationship with a significance level of less than .0000.

If the researcher prefers an exact computation, he can include the option, “exact” in the cci command. That would give a Fisher’s Exact test instead of a χ^2 test. The probability level given with the exact test indicates the proportion of cases with distributions as extreme or more extreme than the one shown in the table.

Case-Control Studies with Several Levels of Exposure

There are frequently situations where there are more than two levels of exposure. Let us suppose that there are k levels of exposure. Often, the exposure occurs naturally at these levels. The levels are grouped according to a natural ordering of the dose-response. Breslow and Day (1980) show how the researcher could employ a 2 by K table to represent the findings (Table 2).

Table 2 A 2 by K Case-Control Analysis					
Exposure level					
	1	2	K	
Cases	A_1	A_2	A_k	N_0
Controls	C_1	C_2	C_k	N_1
Totals	M_1	M_2	M_k	T

In general, one exposure level, for example, level 1, is chosen as the baseline for the analysis. Each of the other exposure levels is compared with the baseline exposure level. From

each of these comparisons, the researcher obtains relative risks RR_1, RR_2, \dots, RR_k along with confidence intervals for each of these levels. He also obtains a test of the hypotheses that all of their values are simultaneously equal to unity. The null hypothesis is that there is homogeneity of all the odds. If the null hypothesis were true, then the expected value of the cells would be

$$E(a_k) = \frac{M_k N_1}{T}. \quad (13)$$

The variances of the cells are given by

$$Var(a_k) = \frac{M_k(T - M_k)N_1N_0}{T^2(T - 1)}. \quad (14)$$

The test statistic testing the homogeneity of k proportions (Breslow and Day, 1980) is given in Eq.12 is a χ^2 test for $k-1$ degrees of freedom.

The case-control data (Breslow and Day, 1980) from the Ille-et-Villaine study of esophageal cancer and the ordinal exposure to alcohol are entered according to Figure 8.

The screenshot shows the Stata Editor window with a table titled "Alcohol[9] =". The table has 8 rows and 6 columns: "Alcohol", "Case", "Freq", "agegrp", and "tobacco". The data is as follows:

	Alcohol	Case	Freq	agegrp	tobacco
1	1	1	2	4	1
2	1	0	47	4	1
3	2	1	9	4	1
4	2	0	31	4	1
5	3	1	9	4	1
6	3	0	9	4	1
7	4	1	5	4	1
8	4	0	5	4	1

Figure 8 Case-Control Data for 2 by K Tabular Input

Once that is done, the STATA command to perform the analysis is typed into the command window:

tabodds Case Alcohol [fweight=Freq]

```
. tabodds Case Alcohol [fweight=Freq]
```

Alcohol	cases	controls	odds	[95% Conf. Interval]	
0-39g	2	47	0.04255	0.01034	0.17518
40-79g	9	31	0.29032	0.13822	0.60979
80-119g	9	9	1.00000	0.39695	2.51919
120+g	5	5	1.00000	0.28950	3.45420

Test of homogeneity (equal odds): chi2(3) = 22.22
Pr>chi2 = 0.0001

Score test for trend of odds: chi2(1) = 20.85
Pr>chi2 = 0.0000

Figure 9 Output from Case-Control Study with Several Exposure Levels

The results appear in the output window, shown in Figure 9. Because the table is a 4 by 2, the test for equality of the odds ratio is a χ^2 with $(K-1) \times (2-1)$ degrees of freedom. In this instance, that means 3 degrees of freedom. Because the test for homogeneity of the odds is significant at the .0001 level, there is indication that there is a difference among the exposure levels. The Score test for trend provides indication that there is a dose-response relationship.

If we were to add the graph option to the command, as follows, one can graph the odds ratio against the levels of alcohol consumption in Figure10 below.

tabodds Case Alcohol [fweight=Freq], graph

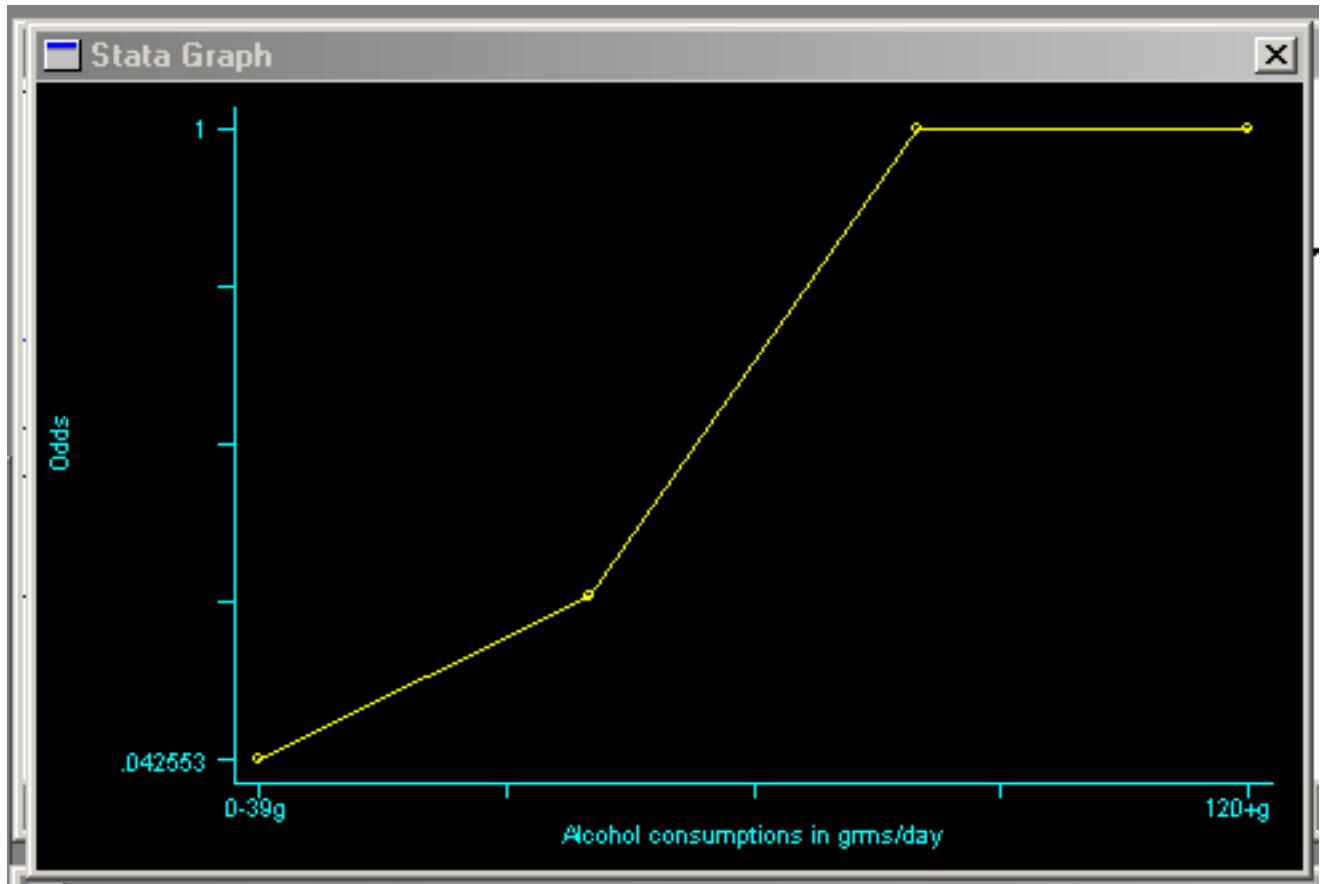


Figure 10 Graph of Odds Ratios by Alcohol Consumption measured in grams per day

Stratified Case-Control studies

STATA permits the researcher to perform stratified case-control studies. When the data are stratified by another variable, a test of homogeneity of the odds ratio can be performed. When the Rothman 1982 data cited in the STATA 7 Reference Guide, A-G are input in tabular form, the command to generate the stratified case-control analysis is

cc case exposed [freq=pop], by (age)

The output appears in Figure 11:

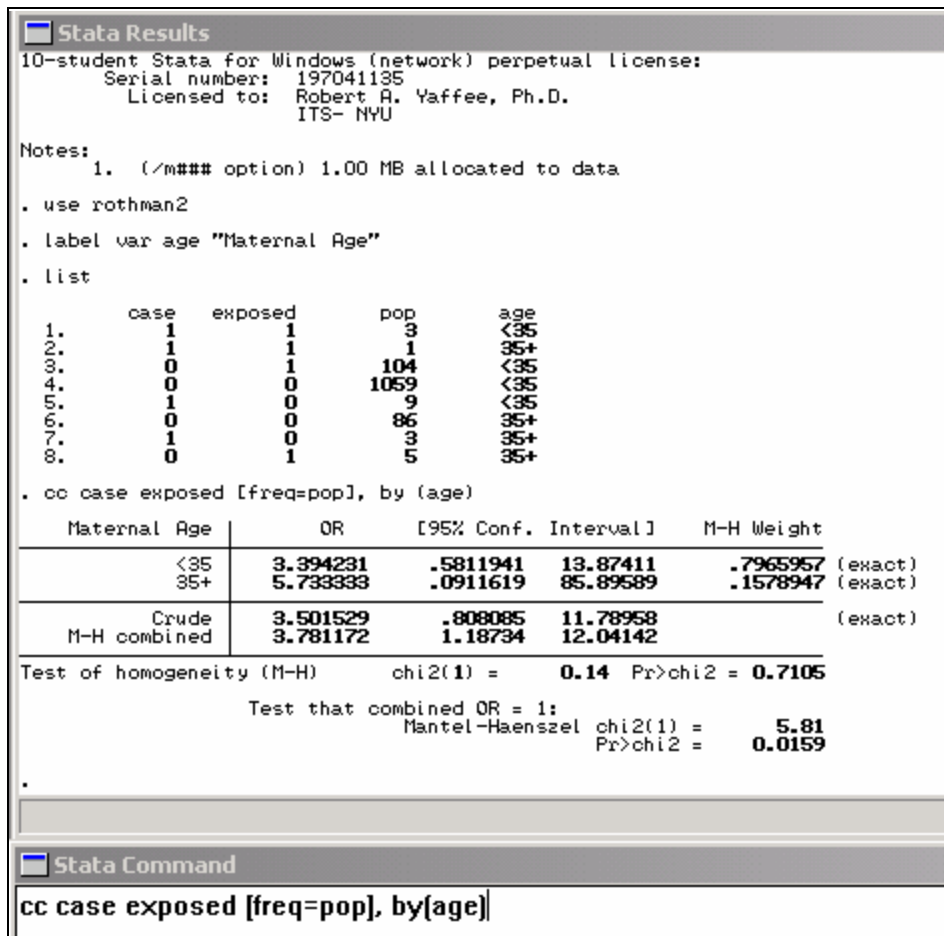


Figure 11 A Stratified Case-Control Analysis

In this case, the stratification is by age and the odds ratio, along with their 95% confidence intervals, of each age stratum is computed. The Mantel-Haenszel Weight (Eq. 7) is reported. In this case the test of homogeneity of strata reveals a statistically insignificant difference. The strata are essentially the same, but when they are combined, the test that the combined odds ratio equals 1 is rejected at the .05 level of significance.

Limitations of Simple Tabular Analysis

Simple two dimensional tables, revealing the relationship between cases and exposure, represent the apparent association between two variables. Analysis with 2 by 2 or 2 by k tables has limitations. Simplification of a complex analysis with a 2 dimensional table may obscure important relationships. Simpson's Paradox is an example where the failure to control for a third intervening or antecedent variable may obscure the real nature of the association.

Problems with the research design may confound the observed results. There must be sufficient **statistical power** for the asymptotic tests to be effective. Without that power, it behooves the analyst to employ the exact tests to minimize the possibility of the type I error (a

false positive significance where a rejection of the null hypothesis should have been accepted). Low statistical power does not protect him against the type II (are false negative significance where the null hypothesis accepted by mistake) or type III error (where he did not ask the correct questions in the first place).

This observed association can be plagued with **selection bias**. Particularly in case-control studies, where the criteria are used for selecting the participants in the groups are not strictly comparable, bias can be introduced into the findings. This usually takes place where incomplete criteria for selection of the groups are employed. Then some of the omitted relevant criteria can confound the findings.

Other biases can come from **interviewer bias, observer bias, or recall bias**. These sources of distortion can produce skews in the results recorded. Observer bias can lead to **misclassification or misspecification bias**. The errors in measurement of exposure or disease can come from lack of sensitivity or specificity in the screening instrument. This number of false positives or false negatives biases the classification or specification process (Breslow and Day, 1980; Ingelfinger et al. 1994 ; Selvin, 1996). Recall bias or **carryover effects** may distort the findings unless they are controlled for by specific research (for example, crossover) designs. Recording errors may stem from typographical mistakes in data input. Data cleaning in the form of multiple inputs and cross-checking can be used to elimination data input errors. In the data collection and input process, all of these sources of erroneous information must be guarded against.

This observed association can be **confounded** by variables associated with the disease and the exposure that are left out of the analysis. These omitted variables may be intervening variables or antecedent variables. They may enhance or suppress the observed relationship between the exposure and the disease. When these omitted variables are unknown and unmeasured, they will confound any relationship examined. The two-dimensional table may reveal the main effects and interaction between two variables. If other-- for example, ecological variables--that are related to both disease and exposure are omitted from the analysis, the table will be confounded by **ecological bias** and therefore cannot properly represent the causal modeling among the related variables.

Epidemiologists can employ **randomization, restriction, matching, and statistical adjustment** to control for confounding problems. Randomized assignment to groups equally distributes the confounding factors in such a way, apart from sampling variation, they should cancel one another out, thereby overcoming selection bias. If the sample size is large enough, the randomization tends to have this effect; if the sample size is too small, an unhappy randomization can perchance unfairly distribute the confounding factors (Breslow and Day, 1980).

Epidemiologists can restrict those participating in the cohort study to avoid potentially confounding factors. The use of exclusion criteria at the beginning of a clinical trial can minimize the probability of biasing the findings. For example, in an Alzheimer's disease clinical trial, researchers would exclude all patients with schizophrenia at the beginning of the clinical trial to avoid confounding their findings. The relationship between exposure to alcohol

consumption and esophageal cancer could be confounded if smokers were not excluded from the analysis. Study restrictions that exclude participants with confounding factors is a standard part of clinical trial, cohort study, or case-control research design.

Matching is sometimes used in case-control studies to overcome problems of confounding. The subjects in the different case and control groups are matched on potentially confounding patient characteristics. Once the data on these matching variables are recorded for the participants, the participants in the control group may be selected according to them. If a many-to-one matching is used, a larger control group may be assembled, thereby enhancing the power of the analysis.

Confounding can be tested and controlled for when the variables are known and included in more complex models. When the variables are known, they may be measured and added into more complex families of regression or survival models. Under these circumstances, there can be statistical adjustment of the confounding covariates that controls for their effects.

References

Breslow, N. E. & Day, N.E.(1980). Statistical Methods in Cancer Research: Vol 1-The Analysis of Case-Control Studies, International Agency for Research on Cancer: Lyon, pp. 1-40, 73-78, 84-115, 122-157, 280-289, 349-351.

Breslow, N. E. & Day, N.E.(1987). Statistical Methods in Cancer Research: Vol II-The Design and Analysis of Cohort Studies. International Agency for Research on Cancer: Lyon, pp.21, 65, 108-109, 336-344, 363-365.

Ingelfinger, J.A., Mosteller, F., Thibodeau, L.A. & Ware, J.H. (1994). Biostatistics in Clinical Medicine. 3rd ed. New York: McGraw Hill, pp. 323-328.

Mausner, J.S. & Kramer, S. (1985). Epidemiology: An Introductory Text. Philadelphia: W.B. Saunders Co., pp.43-64, 312-323.

Selvin, S. (1996). Statistical Analysis of Epidemiologic Data. New York: Oxford University Press, pp. 36, 93-96.

STATA Reference Manual Release 7 Reference A-G, College Station Texas: Stata Press, pp. 455, 466.