Statistics 1: Confidence intervals for proportions; stratified analysis; matched analysis

Obtaining confidence limits for simple proportions

Simple ratios also known as 'binomial proportions such as 69/135 are important in statistics. Sometimes one may wish to know the confidence limits associate with such proportions. In Epi Info 7, just doing the Frequency command on a variable also gives the confidence intervals also known as confidence limits.

1. From the Epi Info main interface select **Classic** (or click the Classic button)
2. Read the project **'Sample prj'** from ***C:\Users\<username>\Epi Info 7\Projects\Sample*** where <username> is the user name on your computer. Open **Smoke**. There should be 337 records.
3. From the Classic Analysis Command Tree, click Statistics > Frequencies. The FREQ dialog box opens. From the Frequency of drop-down list, select **SMOKE**.
4. Click **OK**. Results appear in the Output window.
   - For smoke =1 (Yes) the 95% confidence limits is 20.2% to 29.7%
   - For smoke =2 (No) the 95 % confidence limits is 70.4% to 79.8%

Smoke =1 for 83 or 24.6% of the 337 respondents. If the true frequency in the population is in fact 24.6% and large numbers of samples were taken from the same population, 95% of the results would be expected to fall within the limits 20.2% to 29.7%. These confidence limits might be reported to give an idea of the precision of a simple random survey.

Carry out similar exercise for marital status (**MARITAL**) and see what happens:

Stratified data analysis

A useful technique for analysis in epidemiology is "stratification," meaning that the data are broken into groups or strata based on one or more variables other than the two being used in 2x2 tables.

In this situation there is an "exposure" variable of two levels (e.g., exposed vs. not exposed), an outcome variable with two levels (e.g., the person had the disease or outcome of interest or they did not), and a stratifying variable, which may have two or more levels.

Stratification is performed to investigate whether a stratifying variable is an effect modifier of the exposure-disease relation, a confounder of the relation, or neither. The output provided with stratification includes a table, parameter estimation, and statistics for each level of the stratifying variable; after all the stratum-specific results are provided, the *summary* or *pooled* results are provided along with tests for interaction.

The Evans County cohort heart disease study involves a 7-yr follow of white males raised level of catecholamine (**CAT**) and development of coronary heart disease (**CHD**). The CAT levels have been broken down as high and low and disease outcome has been classified as CHD present or absent.

1. Read the project 'Sample prj' from C:\Users\<username>\Epi Info 7\Projects\Sample\ where <username> is the user name on your computer. Open **EvansCounty**. There should be 609 records.
2. From the Classic Analysis Command Tree, click Statistics > Tables. The TABLES dialog box

opens. From the Exposure Variable drop-down list, select CAT. From the Outcome Variable drop-down list select CHD. *Stratify by*...AGEG1. Click OK.

The stratifying variable AGEG1 is age recoded as "Yes" for those ≥ 55 years of age and "No" for those < 55 years of age. The format of the output is a 2x2 table of CAT by CHD at each level of the stratifying variable, followed by information that summarizes the stratified data. The first stratification table shown in the Evans County stratification example is the relation between CAT and CHD *among younger individuals only* (AGEG1 ="No").

If any stratum has a zero margin, all individuals from that stratum are excluded from the summary analyses. In the Evans County Data, there were 609 individuals in the dataset, and no stratum had a zero marginal value. Therefore, all individuals in the dataset have been used in the summary analyses. If analysis is to be performed on a stratum with a zero marginal sum, that stratum can be collapsed (combined) with an adjacent stratum.

The risk ratio, risk difference, and odds ratios are calculated along with confidence intervals. For each of these parameters, both the "crude" and "adjusted" values are provided. "Crude" values are those in which the stratification is ignored. For example, compare the "Crude" RR, RD, and OR from the Evans County stratification example with the results shown for the Evans County CAT/CHD example, which was based on the TABLES CAT CHD command (i.e., not stratifying on AGEG1). They are exactly the same. In general, the crude parameter estimates will be the same as the parameter estimates from a TABLES command without the stratifying variable when: 1) There are no individuals with known exposure and disease levels but with a missing value for the stratifying variable(s); and 2) there were no strata with a zero margin.

Interpretation of Parameter Estimates and Statistical Tests for Interaction in Stratified 2x2 Tables:

The main reason for stratifying data is to determine whether the stratifying variable modifies or confounds the exposure-disease relationship. The word "modifies" means that the stratifying variable is an "effect modifier" of the exposure-disease relationship or, stated another way, there is an "interaction" between the stratifying variable and the exposure-disease relationship. If there is no interaction, then the issue of confounding is assessed.

1. To determine if there is interaction, look at the p-value for the test for interaction.
   - For the odds ratio, the p-value for the test for interaction (**Breslow-Day test for Odds Ratio**) is 0.81; therefore, there does not appear to be significant interaction.
   - When there is no interaction, the stratum-specific odds ratios will be similar. In the Evans County stratification example, the odds ratio (cross product) for the first stratum was 2.45 and for the second stratum 2.08. When there is a statistically significant interaction, the stratum-specific odds ratios will be different from one another.
2. In the Evans County stratification example, the interaction p-value was not significant; therefore, the next issue is whether the stratifying variable confounds the exposure-disease relationship. To assess confounding, the crude odds ratio is compared with an adjusted odds ratio.
   - For example, compare the crude odds ratio (cross product) with the Mantel-Haenszel adjusted odds ratio [OR(MH)]. The crude OR is 2.86 and the adjusted OR is 2.15. There is no statistical test for confounding; the analyst must choose between the adjusted OR versus the crude OR. Some investigators may choose an arbitrary rule for

deciding whether the level of confounding is "important" by comparing the crude and adjusted parameters. For example, the decision may be: if the crude and adjusted parameters differ by more than 5% or 10%, the stratifying variable will be considered a confounder. If the stratifying variable does not modify nor confound an exposure-disease relationship, then it could be ignored in any further analyses.

3. In the Evans County stratification example, there is large difference between the crude and adjusted odds ratios. Therefore, it appears that age (evaluated as "young" vs. "old") appears to confound the CAT-CHD relationship.
4. The steps for evaluating the role of the stratifying variable are:
   - Is there interaction? If there is interaction, do *not* use the adjusted or the crude OR. Instead, present the ORs from each stratum.
   - If there is **no** interaction: Is there confounding? If the crude OR and adjusted OR are similar, then there is no need to stratify on the variable. If the crude OR and adjusted OR are different, use the adjusted OR because it is more valid than the crude OR.

Matched Pair Case-Control Analysis

The MATCH command from older versions of Epi Info used with matched case-control data is not available in Epi Info 7. In its place is the Matched Pair Case-Control Analysis which is accessible in the Visual Dashboard. With a matched case-control study, it is not generally correct to perform the analyses using 2x2 tables as described above, since this ignores the matching.

The matched pair case-control gadget on the Epi Info 7 Visual Dashboard performs a matched analysis of the specified exposure and outcome variables, which are assumed to be yes/no variables. A table is produced with cells containing the number of match groups showing the combination of positive exposures and positive outcomes shown in the margins. The output table produced by the command is like the summary table produced by TABLES.

The Pair Group ID variable is used for stratification. The matched pair case-control analysis performs a Mantel-Haenszel stratified analysis using the Pair Group ID as the stratifier. Matched groups can have several controls per case, or even more than one case matched to several controls. However, Epi Info 7 currently only does analysis for cases with only 1 control (1:1 matching).

Example

A matched case-control study was performed on twins who served in the military. Individuals with a history of drug abuse were matched with their twin. This would result in matching on age and other general factors. The primary exposure was service in an area of active combat.

Each case and its control were assigned a unique ID. The outcome variable was whether the individual was a "case" i.e., used drugs and usually coded as "Y" or 1 or a "control" (usually coded as "N" or 2). The exposure variable was whether the individual had the exposure (served in an area of active combat, usually coded as "Y" or 1 for those with exposure and "N" or 2 for those without.

1. From the Epi Info main menu select **Tools > Analyze Data > Visual Dashboard** or click the Visual Dashboard button
2. Click on the **Set Data Source** button or arrow. The **Select Data Source** dialog opens. From the Database Type dropdown list change the database type to Microsoft Excel 97-2003 Workbook (.xls)

3. Using the ellipsis (...) button beside the **Data Source** field browse to the downloaded *Course Projects* and choose the Microsoft Excel file **'twins2.xls'**. Select **Sheet1$**. After processing has completed there should be 68 records.
4. Right click anywhere on the Dashboard canvas, select **Add Analysis gadget > Matched pair case-control**. The Matched Pair Case-Control properties dialog opens.
5. From the **Exposure** dropdown list select **service**. From the **Case/Control** dropdown select **drug**. Under **Pair Group ID**, select **ID**. Click **Run**.

Interpretation:

The risk of drug use among military men who served in an active combat zone is 2.8 times (OR MH) greater than the risk among those who did not serve, and that we are 95% of the time confident that the true risk ratio is captured between 1.12 and 7.19 confidence limits which do not include 1.00 and a P-< 0.05 (MH-uncorrected).

If we did an *incorrect* analysis of matched case-control data using the TABLES command, which is essentially an unmatched analysis, the odds ratio would have been 3.8 (95% CI 1.40, 10.48).

While the conclusions would be the same — those military personnel who served in an active combat zone have a dramatically greater risk of drug use than those who did not — the analysis ignores the matching and thus overstates the OR. If the odds ratio is similar in the matched and unmatched analyses, the confidence intervals may also differ between matched and unmatched analyses.