Module 0: Materials

Math competency: Quantitative public health statistics courses require mathematical competency at the pre-algebra level. Topics you need to know include: order of mathematical operations, fractions, decimals, square roots, proportions, percents, ratios, significant digits, scientific notation, solving simple equations, checking solutions, negative numbers, word problems, exponents and logarithms, graphing a line and identifying its slope and intercept, and basic probability. A math self-test will be completed before the semester begins. Please see this handout for additional information: http://www.sjsu.edu/biostat/MathSelfTest.pdf

Procedure notebook: Purchase a three-ring binder in which to store notes. Use this notebook to compile definitions, techniques, computer commands, formulas and interpretative notes. The intention is to create your own data analysis guide. Everything in the notebook should be of your own creation. Please do *not* include photocopies of materials from the text or websites in this particular notebook. (You may use a separate notebook for that purpose.)

SPSS: SPSS is a commercial software package for statistical analysis. The program is installed in the college computer lab. Campus students must apply for a college computer account to use by clicking here. In addition, SJSU has a site license for SPSS. The HS department will provide copies of the software to all graduate students, upon request.

WinPEPI: WinPEPI stands for Windows Programs for EPIdemiologistis. This article by Abramson (2004) describes use of this program. Print this article. In addition, download the program from http://www.brixtonhealth.com/pepi4windows.html and install it on your PC. If you are a Macintosh user, WinPEPI will not run on your computer unless you use parallel Windows processing (e.g., Bootcamp, VMware Fusion). Before the course begins, make sure you can run WinPEPI on your PC or Mac.

Calculator: You must have a TI-30XIIS calculator, and you must be familiar with its operation (especially its order of operation) and the use of these specific function keys: sqrt, x^2, ^, LN, e^x, DATA and STATVAR.

version: July 2009

Lab 1: Inference About a Proportion

Background: Chapter 16.
Data conditions: one sample; binary response; valid data (no information bias); simple random sample / no selection bias.

1. Large sample inference of *p* (*AIDS-related risk factor*).

This part of the lab consists of exercises **16.1, 16.3, 16.5,** and **16.9** in your text. These analyses address the prevalence of an AIDS/HIV risk factor in the U.S. Data are from Catania et al., 1992. Click the highlighted citation and read the abstract. Take note of the study's features.

Notes for each exercise:

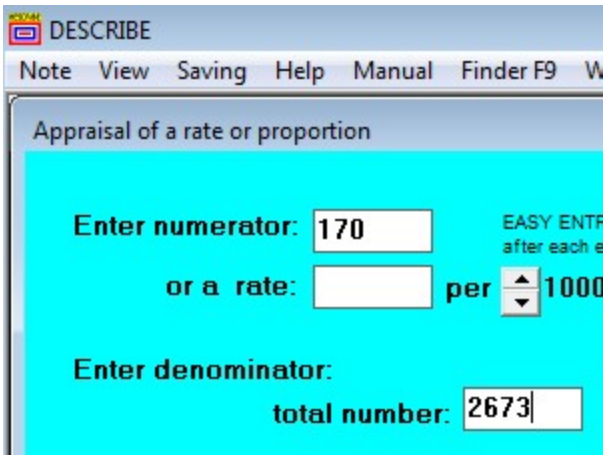
Exercise 16.1 (p. 351): In 16.1.*a*, distinguish between the sample and population; distinguish between the parameter (*p*) and estimate (p-hat). In 16.1.*b* and 16.1.*c*, be clear about the difference between *selection bias* and *information bias,* respectively.

Ex. 16.3 (p. 354): The idea of a sampling distribution is illustrated in Figure 16.3. The binomial distribution is suited for studying random variation of binomial counts and proportions. In large samples, the binomial distribution becomes a Normal distribution ("Normal approximation to the binomial"). Do *not* be concerned about the binomial and Normal equations. Instead, focus on the random number of "successes" in a given sample and how the binomial and Normal approximation are used to address sample-to-sample variation.

Ex. 16.5 (p. 357): Show all hypothesis testing steps. We use these steps when testing statistical hypotheses:

- Step A: Hypothesis statements $H_0$and H_a. Be aware of how the hypotheses statements relate to the research question. This is key to understanding the procedure.
- Step B: Test statistic. The test statistic is dictated by the sample type (SRS) and measurement scale (binary). In this instance, we use a Normal approximation to the binomial (pp. 354 - 5) because outcome is the number of successes in *n* independent trials (binomial) and the sample is large (Normal approximation).
- Step C: *P*-value - Try using "WinPepi > Whatis.exe > P-value" to convert your *z* statistic to a *P*-value.
- Step D: Interpretation of the *P*-value. The *P*-value can be interpreted only in the context of the claim made by H_0. The smaller the *P*-value, the more dubious the claim of H_0. For guidelines on the interpretation of *P*-values, see pp. 181 - 183.

Ex. 16.9 (p. 366): After calculating the confidence interval by hand (formula on p. 363), check your work with *WinPEPI > Describe > Program A*. Enter the number of success in the "Enter numerator" field. Enter the sample size in the "Enter denominator - total number" field:



WinPepi calculates confidence intervals (CI) for *p* using five different methods. *Wilson's score method* (uncorrected) mirrors plus-four hand-calculations. When interpreting the CI, keep in mind that it seeks parameter *p, not* sample proportion p-hat.

The hypothesis test in exercise 16.5, and the CI in exercise 16.9 are bridged by this fact: when the (1-alpha)100% CI excludes the value of *p* specified by the null hypothesis (p_0), then data provide good evidence against the null hypothesis at the alpha level of significance. In execerise 16.5 we tested H_0: $p = .075$. In exercises 16.9 we estimated *p* to be between .055 and .074 with 95% confidence. Thus, we have ruled out a population prevalence of .075 with 95% confidence, corresponding to an alpha level of .05.

2. Small sample inference of *p (Patient Preference).*

This part of lab consists of **exercises 16.2, 16.4, 16.7, 16.8, and 16.11**. Data from a study by Brooks et al. (1998) are used to assess patient preference for one of two medical procedures. *Because of the small size of sample (n = 8), Normal approximation (z) procedures cannot be used.* Instead, exact binomial procedure are used (section 16.4).

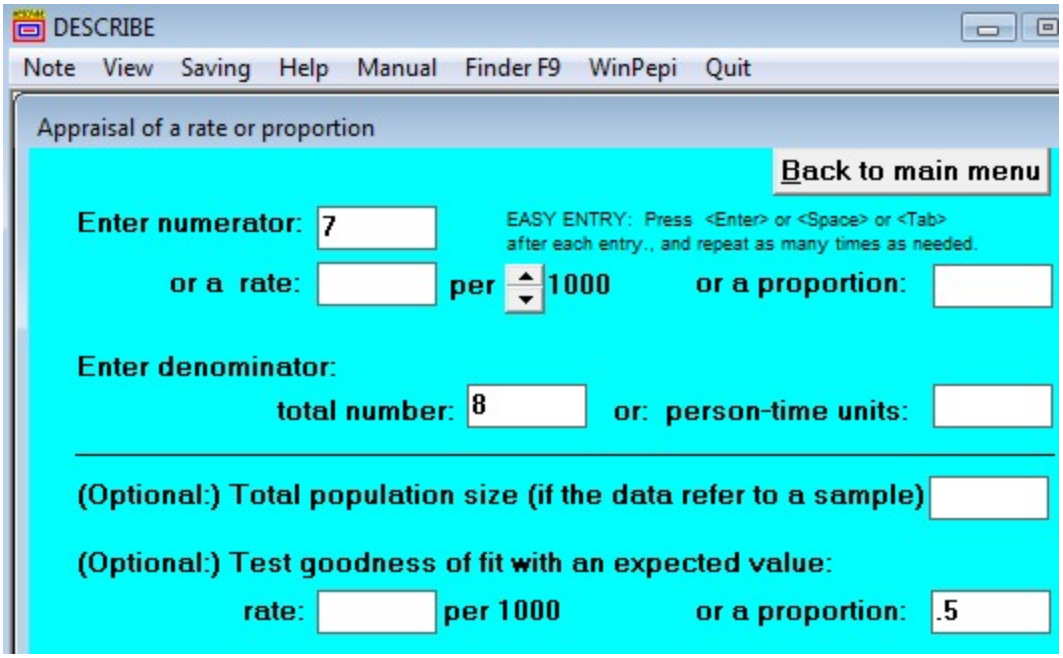
Notes:

Ex. 16.2 (p. 352): In 16.2.*a*, distinguish between the population proportion and sample proportion. The goal of part *b* is to get you to think about the random number of successes in the sample *when* there was equal preference for the two procedures in the population. Under the assumption of equal preference ("no difference in preference"), 4 of the 8 patients would prefer method A and 4 of the 8 patients would prefer method B, *on the average*.

Ex. 16.4 (p. 354): Let preference for method A represent a "success." Because of the small sample, a binomial model must be used to address the random number of sucesses in a given sample. Binomials have two parameters: *n* and *p*. The value of *n* is 8 for our binomial model; the value of *p* is *un*known.

Ex. 16.7 and ex. 16.8 (p. 362): List all steps when conducting a statistical hypothesis test. These are:

- Step A (H0): The null hypothesis for this problem is "no difference" or "equal preference." Because this is a one-sample test of a proportion, this translates to p = .5.
- Step B (Test stat): With exact binomial tests, there is no test statistic *per se*. We merely reiterate the observed count of successes and sample size.
- Step C (P-value): Use WinPepi > Describe.exe > Program A to calculate the *exact binomial P-value*. In the data entry screen, enter the observed number of success in the "Enter numerator" field. Enter the sample size in the "Enter denominator" field. Enter the expected proportion under the null hypothesis in the "Test of goodness of fit with an expected value" field. Your screen should look like this:



Run the program and review the output. Find the section of the output that says "GOODNESS OF FIT WITH EXPECTED PROPORTION." Ex. 16.7 asks for the exact binomial *P*-value. Report the one-sided and two-sided results. Ex. 16.8 asks for the Mid-P exact *P*-value. Both exact tests are used in practice.

- Step D (interpretation): Don't forget to interpret the P-value in the context of the null hypothesis.

Ex. 16.11 (p. 366): Because of the very small sample size, exact binomial confidence intervals should be used. Find the "Exact 95% C.I. (Fisher's)" in the output you produced for exercises 16.7 and 16.8 and report this CI.

3. Sample size requirement to limit margin of error. Complete exercises 16.19 and 16.20 on p. 371. Note the "inputs" and assumptions. Check your calculations with "WinPepi > Describe.exe > K. Sample size > Estimate a Proportion". Did your hand calculations check out? What effect does decreasing the desired margin of error have on the sample size requirements?

Link to notes

Lab 2: Comparing Proportions

Background: Sections 17.1 - 17.5.
Data conditions: Independent groups (binary explanatory variable); binary response variable.

A. Comparing proportions, large samples (WHI trial)

1. Data: Data from the Women's Health Initiative (WHI) trial are used to evaluate the risks of serious health outcomes (e.g., death, heart attack, stroke, breast cancer, and other events) associated with post-menopausal estrogen. Read the Abstract and Introduction of this article: Writing Group for WHI, 2002. Focus on the research question, data conditions, and study design.

2. Incidences, RR, RD: Cross-tabulated results for the combined index outcome (heart attacks, strokes, cancer, and other life threatening events) show:

	Outcome+	Outcome -	Total
Estrogen +	751	7755	8506
Estrogen −	623	7479	8102
Total	1374	15234	16608

a. Calculate the incidence proportions (average risks) of the combined index outcome in each of the groups (pp. 373 - 4). Recall: Incidence proportion = average risk = (no. of onsets) / (no. at risk) or, as formulas, $\text{p-hat}_1 = a_1 / n_1$ and $\text{p-hat}_2 = a_2 / n_2$.

b. Compare the incidences in the form of a ***relative risk* (*RR*)**, i.e., calculate and interpret the *RR* of the index outcome associated with estrogen exposure.

- Review the section on Relative Risks on pp. 389 - 390 in the text.
- To get a relative risk (*RR*), *divide the* risk (incidence proportion) in the exposed group by the risk in the non-exposed group: $\text{RR-hat} = \text{p-hat}_1 / \text{p-hat}_2$
- All intermediate *calculations* should carry at least four significant digits. However, round the results to three *significant* digits just before reporting (eg., RR-hat = x.xx).

c. Compare the incidences in the form of a ***risk difference* (*RD*)**. Risk differences are discussed on p. 375 in the text. The word "difference" simply means "subtraction." To get a risk difference, *subtract* the risk in the non-exposed group from the risk in the exposed group: $\text{RD-hat} = \text{p-hat}_1 - \text{p-hat}_2$. The RR and RD are both *measures of effect.* The *RR* is a measure of relative effect; the *RD* is a measure of absolute effect. See point 5 on p. 390 for a comparison of the these two effect measures.

In this lab, we are going to focus on the *RR*.

3. CI for the RR by hand: The relative risk calculated in 2*b* is the *point estimate* of the RR parameter. The provides the most likely value of the parameter, but provides neither information about other possible values for the RR nor the precision of the estimate. For these purposes, we need a confidence interval (CI) for the RR. Calculate the 95% CI for the *RR by hand using the* method described on p. 391 in the text. *Interpret* your confidence interval, keeping in mind that the CI is trying to locate the ["true"] *RR* parameter.

4. WinPEPI calculation. Check your calculations with "WinPepi > Compare2.exe > Program A". Save or print the output. Find the RATIO estimates on your output (about half way down) and write down the "traditional (log-transformation) method" 95% CI. These CI limits should match your hand calculations.

5. Hypothesis test. It is typical to test the association for statistical significance.

Step A. The null hypothesis can be stated in several ways. One way is H_0: $p_1 = p_2$. An equivalent form is H_0: $RR = 1$. Both claim "no association in the population" between the exposure and disease. Take note of these equivalent ways to state the null hypothesis.

Step B. Several different statistics can be used to test the hypothesis. Page 381 in the text calculates a z_{stat} of 2.66. (You are *not* being asked to calculate this statistic.) The chi-square is an equivalent statistic. There is this simple relationship between chi-square statistics and z statistics for 2-by-2 tables: *the square root of the chi-square statistic is equal to the z statistic; the square of the the z statistic is the chi-square statistic.* Square the z statistic to get the chi-square for this test.

WinPEPI does calculates a chi-square statistics for 2-by-2 tables. Review the WinPEPI output produced in ***3a*** of this lab. Find the chi-square statistics in this output. It will look like this:

```
Pearson's chi-square      =  7.101   P = 0.008  [ 7.7E-3 ]
with Yates's correction =  6.952   P = 0.008  [ 8.4E-3 ]
```

Note that this chi-square statistic comes in regular and continuity-corrected forms. The Pearson's chi-square is the regular chi-square, and the Yates is the continuity corrected chi-square. Focus on the regular (Pearson's) chi-square.

Step C. Note that the chi-square (and z) statistics for this problem produce $P = 0.008$.

Step D. Interpret the result of your hypothesis test. Is the evidence against the null hyopthesis weak or strong? What does this mean in plain language?

B. Further inquiry into the health effects of postmenopausal estrogen

1. Prior studies on post-menopausal estrogen. This part of the lab looks at results from the WHI study in the wider context of what was previously known about the health effects of post-menopausal estrogens. Before the WHI trial, it was widely believed that estrogens had a net beneficial effect on women's health, primarily in terms of cardioprotection. "Potential cardioprotection was based generally on supportive data on lipid levels in intermediate outcome clinical trials, trials in nonhuman primates, and *a large body of observational studies suggesting a 40% to 50% reduction in risk among users of either estrogen alone or, less frequently, combined estrogen and progestin.*[2-5]" (WHI, 2002). What relative risk is associated with a 40% reduction in risk?

Note: The baseline RR is 1. A 40% reduction implies .40 less than 1. See point #2 on page 390 for information on *risk relative to baseline*.

2. Sytematic error = bias. The *RR* for cardiovascular outcomes estimated by the WHI trial was 1.22 (Table 2, p. 326). We may *explain* the discrepancy between earlier observational studies and the results from the WHI trial in terms of *systematic errors (biases)* in the observational studies. The primary way to understand bias in public health research is to consider these three sources of bias:

> **(a) Confounding** - mixing together of effects of the explanatory variable with effects of extraneous "confounding" variables.
> **(b) Information bias** - the mismeasurement or misclassification of variables
> **(c) Selection bias** - the selection of subjects for study in a way that systematically favors a certain outcome.

Propose a plausible explanation as to *how* confounding could cause the observational studies to show a net cardioprotective effect while the WHI trial showed just the opposite.

3. Resolving the discrepancy. Our altered understanding of the health effects of estrogen provides a stunning illustration of how confounding and other forms of bias must be considered when interpreting public health studies. One cannot over-simplify public health research, even when based on published reports: *to do so would devalue objectivity, and devaluing objectivity has real health consequances.*

4. *This part is optional and is covered only with selected groups:* Here's a link to an article by Prentice et al. (2005) that discusses the results from observational and experimental studies on post-menopausal hormone use. Read the abstract and introduction of this article. Your instructor may use this article to discuss specific points of interpretation (varies from semester to semester).

C. Small samples (post-operative exposure & colonic necrosis)

1. Data: Data from a study by Gerstman et al. 1992 are described in this PubMed abstract. Click on the link and read the abstract. Take note of the study's purpose and design.

Data from this study address whether post-operative exposure to a drug called sodium polystrene (brand name Kayexalate) causes colonic necrosis (gangrene of the intestine). Cross-tabulated data are:

	Necrosis+	Necrosis-	Total
Exposure+	2	115	117
Exposure-	0	862	862
Total	2	977	979

Calculate the incidence of necrosis in each group. Comment on these incidences. Does there appear to be a difference in risks?

2. Expected frequencies: We test whether the observed differences in risks is statistically significant. Part A.of this lab used a *z* or chi-square test to test H_0: $p_1 = p_2$. However, z and chi-square tests are based on Normal approximations to the binomial, which should avoided in small samples. Use the method described on p. 423 in the text to determine if a sample is too small to use a z or chi-square proceduer, i.e., check to see whether an expected frequencies in the 2-by-2 table is less than 5. Show your work.

3. Fisher's exact test: Fisher's test is used when Normal approximations (i.e., z and chi-square tests for proportions) cannot be used. We will use WinPEPI to conduct a Fishers. But first we must follow these procedural steps:

a. State the null and alternative hypotheses for testing the proportions for inequality.
b. Fisher's test has no test statistic *per se.* Therefore, we merely report the observed frequency table.
c. Use "WinPepi > Compare2.exe > Program A" to compute Fisher's two-sided *P*-value Interpret this result.

4. Summary. Report the results of this study in narrative form. Address the counts, proportions, and hypothesis testing results using concise and professional language.

Link to notes

[Last edit: 6/19/09]

Lab 3: Naturalistic and Cohort Samples

Background: Text sections 17.6, 18.1 - 18.3.
Data conditions: 2 or more independent groups (purposive cohorts or naturalistic sample), categorical explanatory variable; binary response variable.

A. Chi-Square Distributions

The chi-square probability density function (pdf) is one of the most common probability functions in public health statistics. Study p. 421 in your text for an introduction to chi-square distributions. Note the chi-square shape, degrees of freedom, and relationship to the Standard Normal distribution.

B. 2-by-2 cross-tabulation (prison.sav)

1. Data : A study published Smith et al., 1991 forms the basis of this analysis. Click the highlighted citation to view the article. Read the Introduction and Methods sections of the article. A subset of these data stored in prison.sav. Here's a codebook for the data set:

Variable Information

Variable	Position	Label	Measurement Level	Column Width	Alignment	Print Format	Write Format
hiv	1	HIV serology	Nominal	8	Right	F8	F8
ivdu	2	Intravenous drug use	Scale	8	Right	F1	F1

Variables in the working file

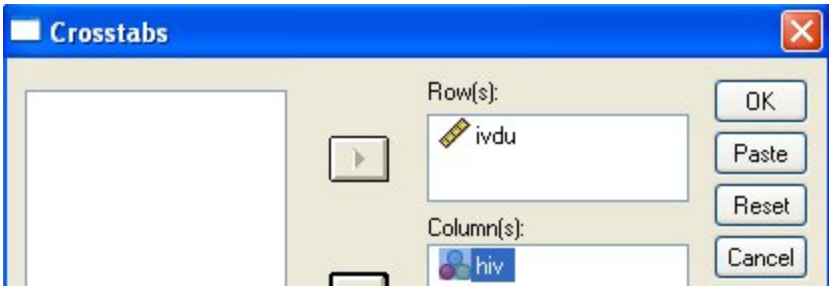
Variable Values

Value		Label
hiv	1	yes
	2	no
ivdu	1	yes
	2	no

Download prison.sav and save it for future use.

2. Get your variables straight:
a. In studying the association between IVDU and HIV, which variables is the explanatory variable? In epidemiologic studies, the explanatory variable is often referred to as the "exposure."
b. Which variable is the response variable? In epidemiologic studies, the response variable is often referred to as the "disease."
c. What measurements scales (quantitative, ordinal, categorical) are used to record these variables?

3. Cross-tab: When working with categorical variable, the first step is to cross-tabulate the data. Use SPSS to cross-tabulate the data via SPSS Analyze > Descriptive Statistics > CrossTab. Put the explanatory variable in the row field and the response variable in the column field.



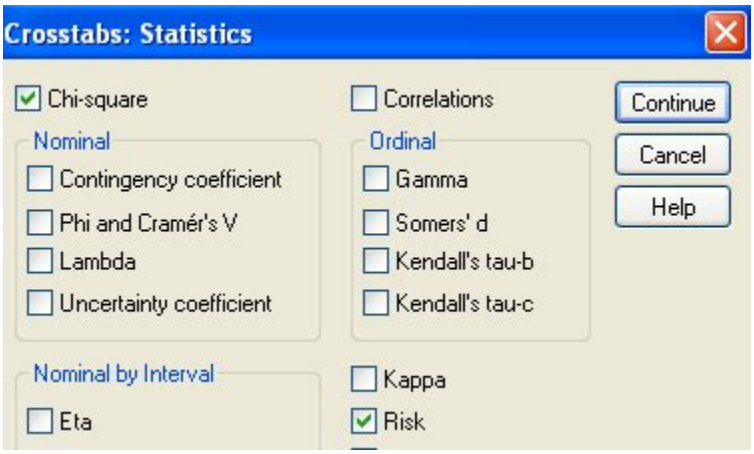
Click Continue > OK and go to the output window. Save the output for further analysis.

4. Prevalence estimates: Next, we calculate the prevalence of the disease in each group.
a. Calculate the prevalence of HIV in the IVDU+ group. Recall: $\text{p-hat}_1 = a_1 / n_1$.
b. Calculate the prevalence of HIV in the IVDU- group. Recall: $\text{p-hat}_2 = a_2 / n_2$.
c. How do these prevalences compare?

5. *RR* estimates: Prevalences are compared in the form of a ratio. The prevalence ratio is referred to as a relative risk (*RR)*.

a. Calculate *RR*-hat and the 95% confidence interval (CI) for *RR* for the association between IVDU and HIV. (Use the same formula as last week). Interpret these statistics.

b. Return to SPSS and click Analyze > Descriptives > CrossTab. Click the Statistics button and check the options for "Chi-square" and "Risk":



Go to the section of output labeled "Risk estimate." [This is a *misnomer:* these statistics are *not* risk estimates; they are *relative* risk (RR) estimates; big difference!] Find the RR estimate and CI for the HIV-positive cohort. These should match your hand-calculations.

Note: When working with already-cross-tabulated data, use "WinPEPI > Compare2 > Program A" to calculate CIs for RRs.

6. Test of the association. We test the association for significance.

Step A: The null hypothesis can be stated in several *equivalent* ways, including:

- H_0: no association between row and column variable in the population
- H_0: $p_1 = p_2$
- H_0: $RR = 1$

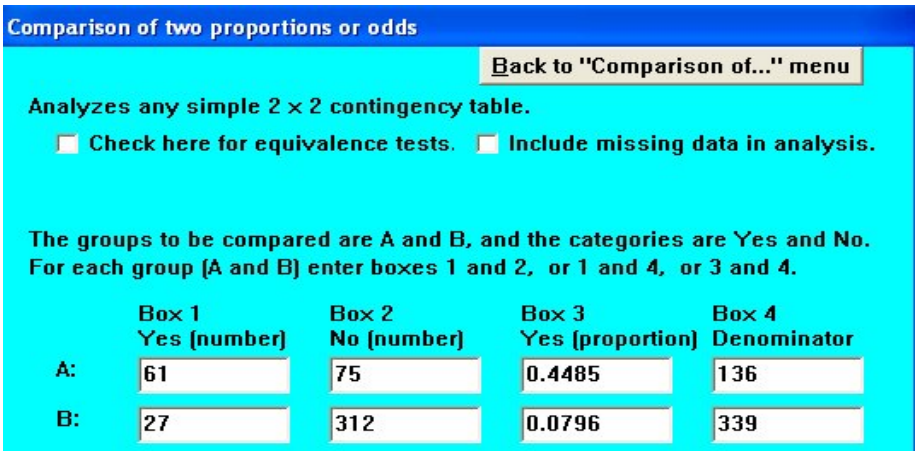
Step B: Several options are available for testing the null hypothesis. Let us Pearson's chi-square statistic (p. 424). The first step in using Pearson's chi-square is to build a table of expected values under the null hypothsis. Then, use the formula on p. 424 to calculate the chi-square statistic. This statistic has (R-1)(C-1) degrees of freedom (p. 425).

Step C: Use Table E as described on p. 425 or WinPEPI > WhatIs > P-value to convert the chi-square statistic and df to a P-value.

Step D: Interpret the *P*-value in the context of the null hypothesis using the guidelines presented in earlier lessons.

b. SPSS . Check your calculations with the chi-square output produced in 5*b*. The Pearson chi-square results should match your hand calculations.

c. WinPEPI. There are times you may be working with cross-tabulate data and need quick and accurate calculations. WinPEPI is an excellent tool for this purpose. Use WinPEPI > Compare2.exe > Program A. to calculate statistics for the current data. You're input screen will look like this:



Find the relative risk and chi-square results on the output screen. The relative risk statistics will be in the section labeled "RATIO [A:B]". The chi-square statistic will be labeled "Pearson's chi-square."

7. Sample size requirements. You are planning a study in this population to determine if race is an independent risk factor for HIV. The study will be restricted to the non-IVDU population. We want to know how many people to study.

The first step in addressing this question is to ascertain what information is needed to determine a reasonable sample size. Page 396 - 400 in the text lists formulas that addresses this question. Note the formula on p. 398. Because of the complex nature of this

formula, software is often used for calculations (see Figure 17.5 on p. 398). Note the following inputs for this calculation:

a) Significance level % (alpha)
b) Power % (1 - beta)
c) Ratio of sample size B:A (r = n_2 / n_1)
d) Proportion in "B" (p_2)
e) To Detect...[three choices] We focus on the detection of a specified [Risk] Ratio A:B (middle choice).

Suppose we want to conduct a chi-square test with an alpha level of .05 (two-sided) and power of .80. Our sample size ratio will be 1:1 (equal group sizes). Assume the prevalence of HIV in the non-exposed group (group B) is .08 and we want to see if the exposed group has *twice* this prevalence. Use WinPEPI > Compare2 > Sample size > Program S1 to determine the sample size requirements for these conditions.

C. R-by-C table (sessmoke.sav)

1. Dataset: Download sessmoke.sav and save it for future use. These data represent a naturalistic sample from a community on smoking and SES. Here is the SPSS codebook for the dataset:

```
[DataSet2] C:\data\datasets\sessmoke.sav
```

Variable Information

Variable	Position	Label	Measurement Level	Column Width	Alignment	Print Format	Write Format
smoke	1	current smoker?	Nominal	8	Right	F8	F8
ses	2	socioeconomic status	Ordinal	8	Right	F8	F8

Variables in the working file

2. SPSS: Open the data file in SPSS. Cross-tabulate the data, placing the explanatory variable in table rows and the response variable in table columns. Click the statistics button and select the "chi-square" option. Run the program, and save the output for review.

3. Prevalences: Calculate the prevalence of smoking within each SES category: p-hat$_i$ = a_i / n_i. You will five prevalences, one for each SES group. To what extent do the five prevalences differ?

4. Test of association: Test the association for statistical significance. Show all hypothesis testing steps (hypothesis statement; tests statistics; P-value; interpretation).
To save time, you may use the chi-square, df, and P-value calculated by SPSS in part 2 of this analysis.

5. Summary. Summarise your results. Taken as a whole, what does Part C of this lab tell you about the prevalence of smoking by SES in this community?

[last edit: 6/19/09]

Link to notes

Lab 5: Stratified Analysis (Confounding and Interaction)

Background: Chapter 19
Data conditions: binary explanatory "exposure," binary response "disease," categorical extraneous "confounding" variables; cohort or case-control sample

A. Selected Points

1. Definition of confounding. Confounding is an *error in inference* due to extraneous factors. Confounding is the most important things to consider when interpreting public health data.

2. Properties of confounders. Confounding occurs when (1) explanatory factor E and potentially confounding factor C are associated; (2) confounding factor C is an independent risk factor for disease D, and (3) confounder C is *not* intermediary in the causal pathway between E and D.

3. Example of confounding. Smoking (confounder C) confounds the relationship between alcohol consumption (E) and lung cancer risk (D) because (1) smoking and alcohol are associated and (2) smoking is an independent risk factor for lung cancer. Alcohol consumers have higher rates of lung cancer, because they are more likely to be smokers, not because they consume alcohol.

4. Mitigating confounding. The text addresses methods that help mitigate confounding on pp. 465 - 466. Study this list! Methods include randomization of the exposure (experimentation), restriction, matching, regression models, and stratification. This Chapter consider a method based on stratification called "the Mantel-Haenszel method".

5. Statistical interaction. *Statistical interaction* occurs when there is heterogeneity in the measure of effect in subgroups. For example, interaction is present when men and women have different relative risks for an exposure-disease relationship. Interaction is also called *effect measure modification*.

B. Crude Analysis

1. Data: Data from Bickel et al., 1975 are used to assess Graduate school admissions at the University of California Berkeley. Click the highlighted text and read the introduction of the article. Briefly, this research considers the application experience of 8,442 male and 4,321 female applicants to UCB Graduate programs in the early 1970s. Assuming men and women who applied for admissions were equally well-qualified, one would expect equal acceptance rates by gender. Although there were over a hundred graduate programs at UC Berkeley at the time of this study, this lab explores the experience of two majors.

Download sexbias2.sav and store it for future use. Here is a codebook for these data:

Variable Information

Variable	Position	Label	Measurement Level	Column Width	Alignment	Print Format	Write Format
sex	1	<none>	Scale	6	Right	F1	F1
major	2	<none>	Nominal	6	Left	A9	A9
accept	3	<none>	Scale	6	Right	F8	F8

Variables in the working file

Variable Values

Value		Label
sex	1	male
	2	female
accept	1	yes
	2	no

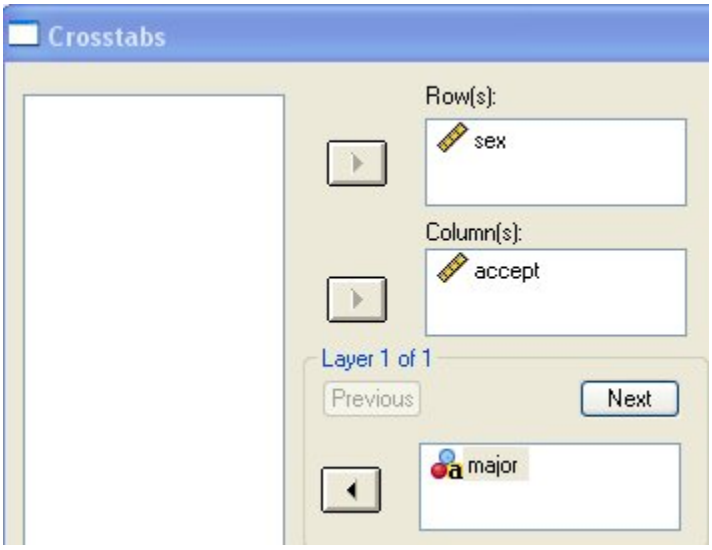
2. ACCEPT by SEX, Crude analysis: Open the data file. Click "Analyze > Descriptive statistics > Crosstabs" and cross-tabulate ACCEPT by SEX. Write down the cross-tabulated results. Calculate the incidence of acceptance for male applicants ($\hat{p}_1$) and female applicants ($\hat{p}_2$).

3. Relative incidence: Calculate the relative incidence (RR-hat) of acceptance for males relative to females. *To what extent* were males more likely to be accepted? *Withhold judgment about cause.*

C. Stratified Analysis

1. Stratified tables. Let's be clear about the variables we are analyzing. What is the name of the explanatory "exposure" variable? What is the name of the response "disease" variable? We wish to address whether the extraneous factor "major applied to" is a confounding variable. What is the name of the extraneous "confounding" variable?

To assess potential confounding, let us stratify the data to create separate analyses at each level of the confounder. **The intent is to create *like-to-like comparisons* in order to mitigate confounding.** To stratify by MAJOR, go to the CROSSTAB dialogue box by clicking Analyze > Descriptive Statistics > Crosstab and add MAJOR as the **layer field.**



Run the program. Print or otherwise save the output.

2. Inspection. You are going to review the strata-specific results using the output you just generated.

a. Calculate the relative risk of acceptance associated with maleness within Major A (i.e., *RR*-hat$_1$). Interpret this result.

b. Calculate the relative risk of acceptance associated with maleness within Major F (i.e., *RR*-hat$_2$). Interpret this result.

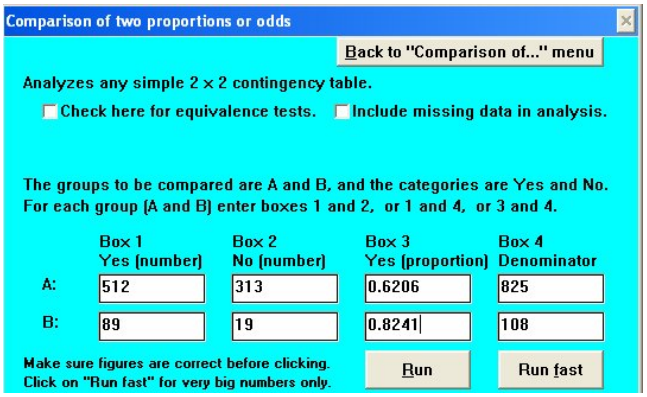
c. Explain why the strata-specific *RR*s showed negative associations between acceptance and maleness while the crude *RR* (calculated in Part B) showed a positive association. Explain the mechanism behind the confounding.

3. Mantel-Haenszel Method

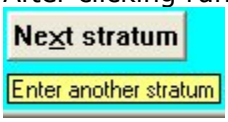
This part of the analysis will use the strata-specific tables to derive a single unconfounded measure of association for ACCEPT and SEX. A Mantel-Haenszel summary RR (RR-hat$_{MH}$) will be used for this purpose. Unfortunately, SPSS does not calculate this statistic. Therefore, let's switch to WinPEPI to complete the analysis.

a. WinPEPI Stratified Tables.

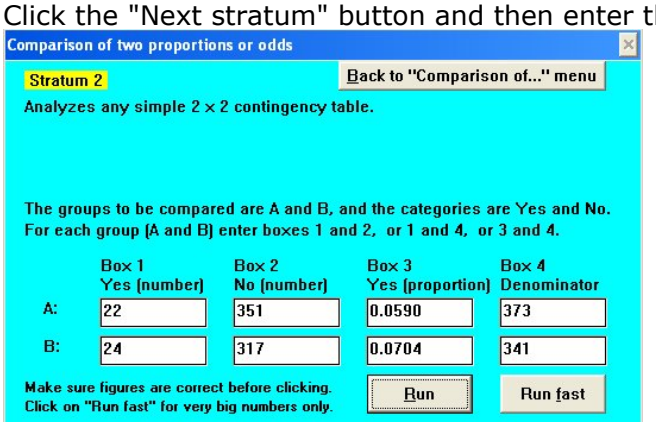
Start WinPepi > Compare2 > Program A. Enter the data for the first strata into the onscreen table. Data for the strata 1 looks like this:



After clicking run, you will see a button on the output screen that looks like this:



Click the "Next stratum" button and then enter the data for the second strata:



After running this second strata, you will see this button:

Run the program for "All strata" and either print or save the results for further analysis.

b. Mantel-Haenszel Test

Near the top of your WinPEPI output, you will see statistics for testing whether the association between the explanatory variable (SEX) and response variable (ACCEPT) is statistically significant after controlling for the confounding factor (MAJOR).

- Step A: The null hypothesis is H_0: $RR_{M\text{-}H} = 1$. That is, we are testing whether the Mantel-Haenszel summary (adjusted) relative risk differs significantly from 1 (indicating a significant association).
- Step B: Use Find the Mantel-Haenszel chi-square statistic and its degrees of freedom in the output. Report these statistics are step B.
- Step C: Report the *P*-value associated with the M-H statistic.
- Step D: To what extent is the evidence against the null hypothesis statistically significant?

c. Mantel-Haenszel Summary RR

Further down in the output, you will see the Mantel Haenszel summary relative risk. This will be labeled "Mantel-Haenszel estimator of ratio." This is a special weighted average of strata-specific RRs and is thus adjusted for the stratification factor MAJOR. Report the M-H summary RR. To what extent are males *less* likely to be accepted after controlling for MAJOR?

4. Interaction

a. Inspection for interaction. List RR-hat1 and RR-hat2. Do these estimates seem homogeneous (interaction absent), or are they significantly heterogeneous (interaction present)? Essentially, we are weighing the evidence to see if the *effect of gender* differs by MAJOR.

b. Test for interaction. Interaction statistics differ for different types of measures of association. We have tests for interaction for the risk difference, and a different test for interaction in the risk ratio. We are focusing on risk ratios to help determine if the heterogeneity in the strata-specific RRs is statistically significant.

- Step A: The null hypothesis is H_0: $RR_1 = RR_2$. This hypothesis states that strata-specific RRs in the population are homogeneous, i.e., no interaction is present.
- Step B: The data is used to assess the evidence against this null hypothesis. Find the heterogeneity chi-square under for the "ratio estimator" in your output. You will find this in the output section that says RATIO OF PROPORTIONS. Report this heterogeneity chi-square and its *df*. Note: The interaction statistic is similar to but differs slightly from the *ad hoc* method in the text; it is suitable for our purpose.
- Step C. Report the *P*-value for the test. Interpret this result. Is there significant evidence against the null hypothesis of "no interaction?" How does this evidence relate to the inspection of the strata-specific RRs you completed in part *a* of this problem?

Link to notes [version 6/23/09]

Lab 6: Correlation and Regression

Background: Chapter 14
Data conditions: quantitative explanatory (independent) variable, quantitative response (depending) variable, linearity. Inferential methods require additional assumptions, e.g., "L.I.N.E."

Part A. Linear Regression

1. Data: Data from an historically important geographic (ecological) study on smoking and lung cancer are:

Country	CIG1930	LUNGCA
USA	1300	20
GrBrit	1100	46
Finland	1100	35
Switzerland	510	25
Canada	500	15
Holland	490	24
Australia	480	18
Denmark	380	17
Sweden	300	11
Norway	250	9
Iceland	230	6

The variable CIG1930 represents per capita cigarette consumption in the year 1930. The variable LUNGCA represents the lung cancer mortality rate per 100,000 person-years in the year 1950. Enter these data into an SPSS data file and save the file for future use.

2. Scatterplot. Use SPSS to create a scatter plot of the relationship between CIG1930 and LUNGCA.
a. Get your variables straight. In regression problems, the explanatory variable is called the independent variable. Which variable is the *independent variable* in this analysis? In regression problems, the response variable is called the *dependent variable*. Which is the dependent (response) variable? Are these variables quantitative or categorical?
b. Create a **scatter plot** demonstrating the relationship between CIG1930 and LUNGCA (SPSS > Graph > Scatter > Define). Make sure you put the independent variable on the horizontal axis.
c. Study p. 296 in your text. Then describe the ***form*** and ***direction*** of the relationship observed in the scatterplot. Are there any potential outliers? (Notice that you are not being asked to describe the strength of the relationship; "strength" is difficult to ascertain by eye.)
d. Explain *why* correlation and regression can be used to describe the relationship between CIG1930 and LUNGCA? Write with the goal of exploring ideas.

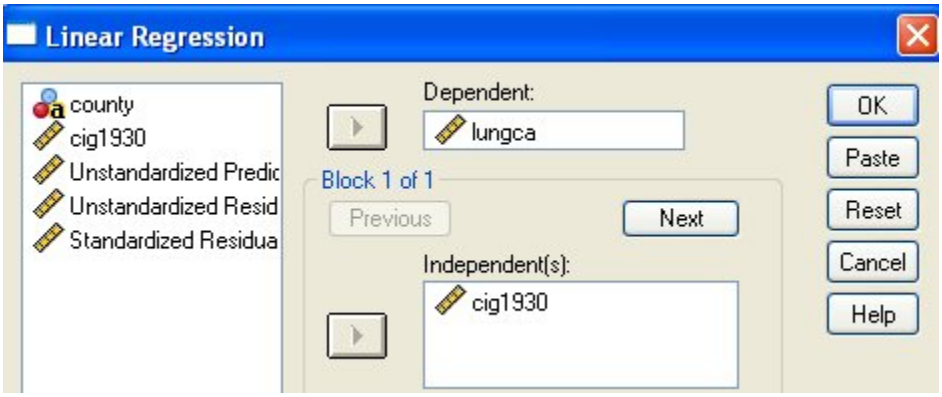
3. Correlation. Pearson's correlation coefficient *r* quantifies the *strength* and *direction* of *linear* associations between quantitative variables.
a. Calculate *r* with SPSS > Analyze > Correlate > Bivariate. The interpretation of correlation coefficients is discussed on pp. 300 -305 in your text. What does *r* tell you about the strength and direction of the association between CIG1930 and LUNGCA?

b. Calculate coefficient of determination r^2. (Merely square the correlation coefficient.) The meaning of this statistic is discussed on p. 300 in your text. What proportion of the variance in LUNGCA is associated ["mathematically explained] by the variation in CIG1930?
c. List the conditions needed to test the correlation coefficient for statistical significance (p. 308).
d. Let us assume all hypothesis testing conditions have been met. Test the correlation coefficient for statistical significance. This hypothesis testing procedure is discussed on pp. 305 - 307.

Step A: List the null hypothesis being tested. Take note of the notation used to denote the parameter being tested and its assumed value under the null hypothesis.
Step B: Do not calculate the test statistic.
Step C: Report the P-value calculated by SPSS.
Step D: Interpret the results of the test procedure.

4. Regression. Review the section on regression in your text, starting on page 311. Regression uses a straight line to model the relationship between the independent variable X (CIG1930) and dependent variable Y (LUNGCA). The slope of the line predicts the **change in Y per unit X**, quantifying the effect of the explanatory factor.

a. Use SPSS to determine the linear regression coefficient that describes the relationship between CIG1930 and LUNGCA. "Regress LUNGCA on CIG1930." The menu choices are SPSS > Analyze > Regression > Linear. In the dialogue box, "Dependent" refers to the response variable and "Independent" to refers to the explanatory variable. Fill in the dialogue box like this:



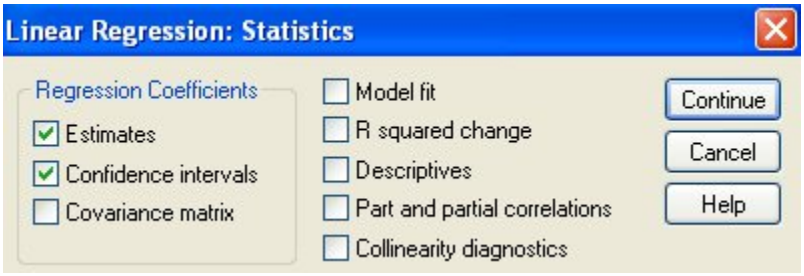
b. Report **slope coefficient *b***. You will find this statistic in the "Coefficients" table at the bottom of the regression output. Look for the unstandardized coefficient associated with CIG1930 variable; this is *b*. Interpret the slope ("predicted change in Y per unit X", p. 313 - 314, note 1) in the context of these data.

The unstandardized coefficient labeled "(Constant)" is intercept estimate *a*. Report intercept coefficient a.

List the regression model in the form LUNGCA = $a + b \cdot$CIG1930 (replacing *a* and *b* with the intercept and slope estimates of the model).

c. Compute the 95% confidence interval for slope parameter "beta" by clicking the statistics button in the Linear Regression dialogue box:

Check "confidence intervals" box in the dialogue box that follows:



The confidence limits for CIG1930 will be reported in the new output. Interpret this interval. As always, confidence interval are interpreted in the context of *parameter* being estimated.

d. Using the regression model for prediction. Predict the lung cancer mortality rate (per 100,000 person-years) in a country with an annual cigarette consumption of 800 cigarettes per capita. This method is described as point 2 on p. 314. Use the formula $\hat{y}_i = a + bx_i$, where $\hat{y}_i$ = is the predicted value, a = the slope estimate, b = the slope estimate, and x_i in this case is given as 800.

e. Residuals.

Recall that the regression model predicts the value of *Y* based on this linear model:

$$\hat{y}_i = a + bx_i$$

where $\hat{y}_i$ is the predicted value for observation *i*, *a* is the intercept estimate, *b* is the slope estimate, and xi is the value of observation *i*.

The *residual* associated with observation *i* is the distance of a data point from its predicted value in the Y direction:

$$\text{residual}_i = (y_i - \hat{y}_i)$$

See Figure 14.9 on page 312 for a visual depiction of the residuals associated with each data point in the data set. Here is a table that depicts residuals. Fill in the missing information in this table, showing the predicted values, residuals, and squared residuals for observations 3 and 5:

i	*X*	*Y*	predicted $\hat{y}_i$	residual $y_i - \hat{y}_i$	residual^2 $(y_i - \hat{y}_i)^2$
1	1300	20	36.453	-16.453	270.704
2	1100	46	31.884	14.116	199.253
3	1100	35			
4	510	25	18.406	6.594	43.475
5	500	15			

6	490	24	17.950	6.050	36.608
7	480	18	17.721	0.279	0.078
8	380	17	15.437	1.563	2.444
9	300	11	13.609	-2.609	6.808
10	250	9	12.467	-3.467	12.020
11	230	6	12.010	-6.010	36.122

The regression coefficients *a* and *b* in our regression model (6.76 and .0228, respectively) were minimize the sum of the square of these residuals ("least squares line"). In turn, the residuals are used to estimate the standard error of the model and other inferential statistics. (You can infer the tedious nature of these calculations.) Finally, we can use the residuals to assess the assumptions of the model (pp. 324). Higher levels of understanding regression models require greater scrutiny of residuals.

SKIP **f. Calculating the confidence interval for beta by hand** ***SKIP*****.** The residuals are used to to calculate the variability of the regression model with a statistic called the standard error of the regression ($s_{Y|x}$). The formula for the standard error of the regression is:

$$s_{Y|x} = \sqrt{\frac{1}{n-2}\sum \text{residuals}^2}$$

$s_{Y|X}$ is then used to calculate the standard error of the slope (SE_b) with this formula:

$$SE_b = \frac{s_{Y|x}}{\sqrt{n-1}\cdot s_X}$$

This standard error has (*n* - 2) degrees of freedom, and the 95% CI for slope parameter "beta" can now be determined with this formula:

$$b \pm (t_{n-2,.975})(SE_b)$$

Calculate the 95% CI for beta by hand.

Part B. Nonlinear relationships

1. Data. Data representing FLUORIDE levels in public water supplies (ppm) and dental CARIE rates per 100 children in 21 North American cities were published in an historicaly important study by Dean et al., 1942. Click the link and read the first page of this article.

Data for the 21 cities in the study are:

OBS	FLUORIDE	CARIES
1	1.9	236
2	2.6	246
3	1.8	252
4	1.2	258
5	1.2	281
6	1.2	303
7	1.3	323
8	0.9	343
9	0.6	412
10	0.5	444
11	0.4	556
12	0.3	652
13	0.0	673
14	0.2	703
15	0.1	706
16	0.0	722
17	0.2	733
18	0.1	772
19	0.0	810
20	0.1	823
21	0.1	37

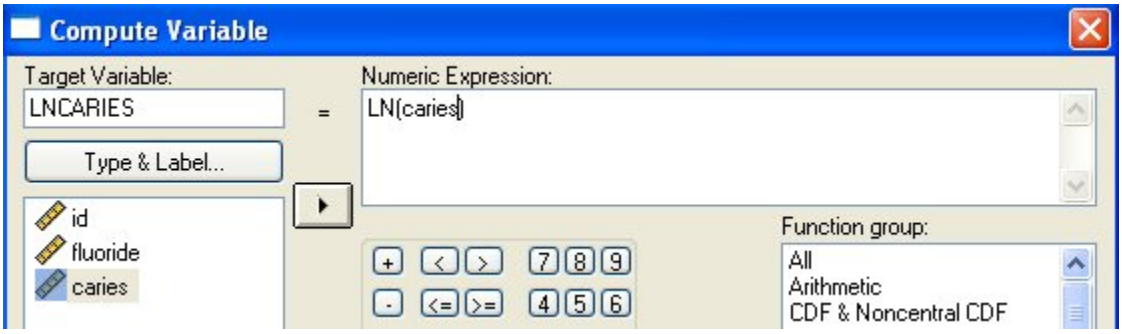
Enter these data into an SPSS file and save the file for future use.

2. Scatterplot. Create a scatter plot of the relationship between FLUORIDE and CARIES using SPSS (Graph > Legacy > Scatter). Make certain the independent variable is on the horizontal axis and the dependent variable is on the vertical axis. Describe the *form* and *direction* of the relationship. Is the association linear? Are there outliers? Can correlation and regression be used to analyze these data? Explain your reasoning.

3. Transformation. Although unmodified regression does not apply to these data, we can still use regression if we first straighten-out the relationship with a mathematical transformation. Let us apply a logarithmic transforms to both the FLUORIDE variable and the CARIES variable to see how this works.

- Make a copy of the data file. Call the copy of the data file *water2.sav*.

- Open water2.sav in SPSS.
- Find the outlier in the data set and delete it.
- From the menu choice Transform > Compute. A dialogue box will appear. Fill in the Target Variable field with the new variable name LNCARIES. Fill the Numeric Expression field with LN(CARIES). Your screen should look like this:



- Click OK and a new variable containing the logarithmic transformed data should appear in your data table.
- Perform a similar logarithmic transform on the FLUORIDE variable. Call this new variable LNFLUOR.
- Plot the transformed variables.
- Review the graph. Is this relationship linear?
- Run correlation and regression analyses on this transformed data. Report salient results, and interpret your findings.

4. Range restriction. An alternative approach for anlaysis is to eliminate the outlier and restrict the range of that portion of data that are *approximately* linear. Let's see how this works:

- Make a copy of the data file. Call the new file water3.sav.
- Open water3.sav.
- Sort the data on FLUORIDE (Data > Sort Case -- select FLUORIDE > OK.)
- In the FLUORIDE and CARIES columns, delete observations with FLUORIDE values greater than 1 ppm.
- Plot the data (Graph > Scatter > Define...). Describe the form of the scatter plot. Is it approximately linear?
- Run a regression on this range of data.

5. Select a model. Which model do you prefer, the model basedf on the untransformed data, the logarithmic transformed data, or the range-restricted data? Explain your reasoning.

Link to notes [Last edited: 6/21/09]