

## PubH 5450 Homework 10 Due Dec. 6

**Exercises from the text:** 7.57, 8.15, 8.38, 9.12.

### Exercises on event related potentials

First copy the files `erpDataEg.dat` and `hw10f01.sas` into your account from my account.

The file `erpDataEg.dat` has event related potential (ERP) measurements for 72 subjects, half of which are infants of diabetic mothers and half of which are healthy infants. The experiment here consisted of attaching 13 EEG electrodes to the scalp of each infant to measure the voltage at each of 13 separate locations in response to a given stimulus. Two stimuli are administered to each infant. One is a picture of the infant's mother, the other stimuli is a picture of a woman the subject has never seen before. The voltages are recorded every 10 milliseconds for about a second and a half following the administration of the stimulus. It is suspected that the pattern of voltage over time depends on whether or not the infant is the child of a diabetic mother. It is thought that the hippocampus of such children is damaged due to fetal iron deficiency, chronic hypoxia and hypoglycemia.

The experiment is repeated dozens of times for each subject. For each trial, we get a sequence of voltage measurements. To simplify the analysis, researchers typically first average all of the trial sequences, then work with the area under the curve (AUC) over a certain time frame for this average ERP. Here we will work with the AUC from 250-400 milliseconds at 3 different locations.

The data file has 2 rows for each subject, one row for each experimental condition (mother versus stranger). The first column indicates which subject the data in that row is for, the next column has AUC for channel 4, the following column has AUC for channel 6, the next is AUC for channel 8, the following column is group membership (1 for infants of diabetic mothers and 0 for controls), and the last column indicates which experimental treatment was administered (0 for mother's face and 1 for stranger's face).

The SAS command file reads in the data, then performs a 2 sample *t*-test to assess if the groups differ in their responses at each of the channels ignoring the experimental condition. You will need to make amendments using SAS techniques from previous homework exercises.

### Questions

1. Use `proc chart` to construct a histogram, and find the mean, median and standard deviation for the response in each channel. Does it appear that the variables are normally distributed? Explain your answer.
2. For each channel, determine if there is a statistically significant (with  $\alpha = 0.05$ ) difference between the two groups ignoring the condition (note that the SAS command file, `hw10f01.sas`, will do this for you without making any changes to it).
3. On a previous homework we learned about the *Bonferroni* method for conducting multiple tests on the same data set (see p. 482 in the text). Use this method (with  $\alpha = 0.05$ ) to assess the statistical significance of observed differences between the groups at the 3 locations.
4. Repeat questions 2 and 3, but now do the analysis for the different conditions separately (use `proc sort` and `by` as we discussed in class).
5. SAS would allow us to do a 2 sample *t*-test to determine if there is a difference between the conditions ignoring the effect of group. Explain why the *p*-values you would find using `proc ttest` (as above) would be incorrect.

### Exercises on the relationship between birth weight and survival

A study was conducted to examine the effect of low birth weight on survival of an infant to one year. The researchers collected all the data that was available from New York City in the year 1974 on birth weight and survival to one year (from hospital records). This data set is taken to be typical of what happens in any given year in New York City during the 1970's.

An infant is said to have low birth weight if it weighs less than or equal to 2500 grams, and is considered normal otherwise. The data is as follows: 618 low birth weight infants died before one year, while 4597 low

birth weight infants survived beyond one year, and 422 normal weight infants died within one year while 67093 normal weight infants survived beyond one year.

### Questions

6. Compute the relative risk for the risk factor low birth weight. Describe in words what this statistic means.
7. Compute the odds ratio for the risk factor low birth weight, and find a 95% confidence interval for this quantity.
8. Use your results from question 7 to test the hypothesis that the odds ratio is 1.0. What do you conclude about the relationship between low birth weight and death before one year?

### SAS techniques for the analysis of categorical data

Previously we learned how to read data in from an external file, now we will see that you can enter the data in the data step inside the SAS command file. Since, in practice, we usually analyze large data sets (i.e. at least 100 observations with several variables), entering the data in the data step is usually not the best method (because it is error prone). An exception to this rule is when we already have the data in the form of a frequency table, as is the case here.

#### Entering data in the SAS command file

As an example, we will enter a table giving the relationship between blood type and some stomach disorders (found in a large retrospective study). The table is

Blood Type	Peptic Ulcer	Gastric Cancer	Controls
O	983	383	2892
A	679	416	2652
B	134	84	570

so we do the following to enter the table. First invoke the data step, give the data a name, then use the input command to name the variables as before. One new thing here is that when you have a qualitative variable (i.e. a variable whose values are not numbers), you instruct SAS that this is the case by ending the variable name with a dollar sign \$. After this, you issue the command `datalines`; to indicate that you are entering raw data. After the `datalines` command, you list the variables in the order in which they are named in the `input` command. Finally end with a semicolon ; on the final line. Here is the data step to enter the previous table into SAS.

```
data tabledata;
  input blood$ cancer$ count;
  datalines;
  O peptic 983
  O gastric 383
  O control 2892
  A peptic 679
  A gastric 416
  A control 2625
  B peptic 134
  B gastric 84
  B control 570
  ;
```

We often want to examine the joint, univariate and conditional distributions of the variables, which are here stomach disorder type (which takes the values peptic ulcer, gastric cancer and none) and blood type (ignoring AB). To this end, use `proc freq` to generate a contingency table and the desired proportions as follows.

```
proc freq data=tabledata;
  tables blood*cancer;
  weight count;
```

Note that we use `weight` to instruct SAS that the variable `count` holds how many observations are in each cell of the table.

We will now learn how to use SAS to compute Pearson's  $\chi^2$  statistic, the relative risk, the odds ratio and how to perform inference for these quantities. To get the desired quantities, first use a data step to set up the table as above. One thing to realize when you do the data step is that SAS constructs tables by ordering the categories on the margins alphabetically by the names you give to the variables, and it puts the first variable to the tables command on the top of the table. This is important to realize when you find relative risks and odds ratios because if you are not wary, you may find that your risky exposure (here low birth weight) has a positive impact on the response (here, surviving to one year). The tip here is to just name the values of the variables intelligently, for example, use `low` for low birth weight infants and `normal` for regular infants (but not `low` and `regular`) and use `dead` or `survive` (but not `dead` or `alive`).

Then we use `proc freq` to compute the desired statistics. Assuming you call the data set with the table you construct in your data step `hw10data`, you refer to the birth weight variable as `brthwt` and the life or death status as `alive`, the syntax is as follows:

```
proc freq data=hw10data;
  tables brthwt*alive / chisq cmh;
  weight count;
run;
```

Use SAS to execute the commands given above (so you must write the data step and use the commands just given).

### Computer Questions

9. To examine the manner in which birth weight influences death, you should first examine either the row or column percentages. Which of these is informative here for the relationship of interest? Report this conditional distribution as it appears on the SAS output.
10. Look at the SAS output and report the value of Pearson's  $\chi^2$  statistic. Does it appear that the relationship is too strong to be due to chance alone (supposing that this year is representative of other years in the 1970's in NYC)?
11. What is the 95% confidence interval for the odds ratio that SAS reports?
12. What is the 95% confidence interval for the relative risk that SAS reports? Can we reject the hypothesis that the relative risk is 10 with  $\alpha = 0.05$  and a 2 sided test?