Laboratory 2  
Homework 2 Answer Key

1. Produce frequency distributions for all the variables of the dataset and check for any inconsistent (or improbable) and out of range values (or outliers). Examine frequency distributions for every variable for the cases and controls separately. If there are any inconsistencies (or improbable values) print the distribution of the variable(s) and mark the inconsistent/improbable value(s) on the printout.

From the output, variables ‘age’ and ‘tobacco’ are within eligible values. Tobacco is within acceptable values.

One case is incorrectly coded as 2. It is a case because the ID value is in the case range. This case needs to be recoded as 1. Please note that IDs in general should be randomly assigned to cases and controls. The systematic number assignment in this dataset is not a general practice.

2. If any of the inconsistencies are correctable based on the information on the study you were given last week, correct them.

There is 1 subject with sex coded as 0 (actual value should be 1). There is 1 subject with status coded as 2 (actual value should be 1). These mistakes are correctable without original questionnaire.

You need the following commands to identify those inconsistent entries.

```plaintext
proc print data=case_control978;  
where sex=0;  
var id sex;  
run;
```

```plaintext
proc print data=case_control978;  
where status=2;  
var id status;  
run;
```

To correct the above mistakes, you need to run the following commands.

```plaintext
data case_control978;  
set case_control978;  
if id=6000 then sex=1;  
if id=6200 then status=1;  
run;
```
3. Are there any inconsistencies in the data that are not correctable without examining the original questionnaire or any outlier(s) (or improbable values) that you would like to verify with the original questionnaire? If yes, indicate which ones.

For the alcohol variable there are 2 abnormally high values (868 and 916). The correct values should be 268 and 216 respectively. These mistakes are not correctable without examining the actual questionnaire.

After checking the correct responses from the questionnaire, correct them with the following commands.

```sas
data case_control978;
set case_control978;
if alcohol=868 then alcohol=268;
if alcohol=916 then alcohol=216;
run;
```

4. Are there any missing values in the data? If yes, on how many subjects and for which variable(s)?

For the tobacco variable there are 7 subjects (2 unknown and 5 did not answer) with missing values (values 8 and 9).

5. Examine the effects of continuous variables on esophageal cancer. Perform independent t-test procedure, and fill in Table 1. Do cases differ from controls with respect to mean age and alcohol consumption? Write a brief interpretation of your results.

Table 1. Mean age and alcohol consumption on esophageal cancer among 978 men

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>t value</th>
<th>DF</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>50.2</td>
<td>14.3</td>
<td></td>
<td>461</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cases</td>
<td>59.9</td>
<td>9.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. alcohol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>44.3</td>
<td>31.9</td>
<td></td>
<td>245</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cases</td>
<td>85.1</td>
<td>48.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cases were older (mean age 59.9, SD 9.5 vs. 50.22, SD 14.3 for controls) and more likely to have consumed alcohol (mean consumption 85.1, SD 48.5 vs. 44.3, SD 31.9) than the controls (Note if you want to use p values to compare these means you should use the p values for unequal variances since the null hypothesis for equal variances is rejected by the F-test at the bottom. However, you should be aware of the limitations of p values in general and in using data description to make inferences, see discussion in class as well as in text).

6. Examine the effects of categorical variables ('agegrp,' 'alcgrp,' and 'tobgrp') on esophageal cancer. Generate 2 x 2 tables, perform $\chi^2$ test, and fill in Table 2. Do cases differ from controls with respect to age, alcohol and tobacco consumptions? Write a brief interpretation of your results. Discuss some drawbacks, particularly in light of what you have read in Rothman and Greenland, of classifying the variables as you have using agegrp, tobgrp, and alcgrp.

Table 2. Effects of age, alcohol and tobacco consumptions on esophageal cancer among 978 men

<table>
<thead>
<tr>
<th></th>
<th>Percentages in Controls</th>
<th>Percentages in Cases</th>
<th>$\chi^2$</th>
<th>DF</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. agegrp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;60 years</td>
<td>69.15</td>
<td>45.50</td>
<td>38.8</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>60+ years</td>
<td>30.85</td>
<td>54.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. alcgrp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;40 gm/day</td>
<td>49.7</td>
<td>14.5</td>
<td>80.9</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>&gt;39 gm/day</td>
<td>50.3</td>
<td>85.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. tobgrp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10 gm/day</td>
<td>57.96</td>
<td>39.39</td>
<td>21.9</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>&gt;9 gm/day</td>
<td>42.04</td>
<td>60.61</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Age, alcohol, and tobacco are all associated with case control status. By examining the percentages that fall into each cell you can also say that cases are older, and more likely to have smoked tobacco and consumed alcohol than controls.
In general, dichotomization loses information about the range of data. The distribution of data is truncated and dose-response relationship may be inhibited. Dichotomization becomes problematic if the cutoff point is based on an arbitrary number. However, in other circumstances, dichotomizing a continuous variable is sensible. For example, a medical diagnosis of high blood pressure versus low blood pressure is made by dichotomizing a continuous blood pressure measurement. Screening scales and psychiatric disorder diagnoses (i.e., depression or no depression) are examples of the usefulness of dichotomization of variables.