ANALYSIS OF CATEGORICAL DATA FROM THE CALIFORNIA DEPARTMENT OF CORRECTIONS

Before starting this lab, you should be familiar with: the difference between categorical and quantitative variables, and how to describe distributions of categorical variables using frequency counts.

About the Data

The dataset you’ll examine is from a study by the California Department of Corrections (CDC) on the effectiveness of prisoner placement, and the likelihood of misconduct while incarcerated. Upon admission to a California prison, an inmate is given a questionnaire. The score on this questionnaire determines which level of security (Level I is the lightest security, Level IV the heaviest) to which an inmate will be assigned. Security is expensive, and so the CDC would like prisoners to be assigned to the lightest security possible. On the other hand, if prisoners are dangerous to themselves or others, or at high risk for escape, then they need to be assigned to a higher level of security. The CDC hired the UCLA Statistical Consulting Center to examine their classification system with the hope of learning whether it "worked." The data you are examining come from that study.

(These data were provided by Professors Richard Berk and Jan de Leeuw, Department of Statistics at UCLA. You’ll find the full study in the article "An Evaluation of California’s Inmate Classification System Using a Generalized Regression Discontinuity Design." Journal of the American Statistical Association, December 1999, Vol. 94 (448): 1045-52.)

Load this dataset into Stata

```
. use http://www.stat.ucla.edu/labs/datasets/prison.dta
```

In this dataset you should find the following variables:
violation: 
Inmates who "misbehave" one or more times, are recorded as having a misconduct violation. In theory, a high-level of security should curtail a convict's ability to commit a violation. Whereas other convicts are placed in low-level security because it is believed they are less likely to commit a misconduct violation. A "1" represents at least one misconduct violation, while a "0" represents no misconduct violations.

score: 
This represents the score from the questionnaire, which can range from 0 to 79. High scores are believed to predict violent behavior and therefore deserve higher security.

strike2: 
Inmates who are serving time for their second strike under California's "3-Strike Law" are given a "1", all other inmates are given a "0".

strike3: 
Inmates who are serving time for their third strike under California's "3-Strike Law" are given a "1", all other inmates are given a "0".

level4: 
Inmates who are classified to security level IV are given a "1", while those inmates who are not in security level IV are given a "0".

Variables "violation", "strike2", "strike3", and "level4" are categorical variables, while score is a numerical variable. Caution: Level IV security level is not another strike under the 3-Strike Law.

Constructing and Interpreting Two-Way Tables
In order for us to evaluate the categorical variables in the dataset we will use two-way tables. Type:

```
. tabulate violation level4
```
Interpreting a Two-Way Table

The chart that you see shows the frequencies for the categories for the inmates who committed violations compared with whether they are placed in level IV security. The values in the first row represent the number of inmates who have committed a misconduct violation. For example, 2299 inmates have not committed a misconduct violation (violation="0") and are not in level IV security (level4="0"), while 486 inmates have not committed a violation (violation="0") and are placed in level IV security (level4="1"). There are 2785 inmates in the dataset who have not committed a misconduct violation.

Question 1: Interpret the values in the second row.

Question 2: How many inmates in the dataset have committed violations?

Now we will consider the row relative frequencies. Type:
\[ \text{. tabulate violation level4, row} \]

In addition to the counts from before, we now have row percentages. Let's consider the first column of relative frequencies. The percentages can be interpreted as follows: Out of the inmates who have not committed violations while imprisoned, 82.55% of them are not in level IV security. This percentage is calculated using \((2299/2785)\). 78.28%, calculated \((890/1137)\), of those inmates who have committed violations are not in level 4 security, and 81.31% \((3189/3922)\) of total inmates in the sample are not in level IV security.

Question 3: Interpret the second column of relative frequencies.

Now let's consider the column percentages. Type:
\[ \text{. tabulate violation level4, col} \]
Again, we have the counts from before, but now we have another set of percentages, the relative frequencies for the columns. These percentages can be interpreted as followed: 72.09% (2299/3189) of the inmates not in level IV security have not committed a violation. Of those inmates in level 4 security, 66.30% (486/733) have not committed a violation. Of the total inmates in the sample 71.01% (2785/3955) have not committed a violation.

**Question 4:** Interpret the second row of relative frequencies.

**Question 5:** Are level IV inmates more or less likely to commit a violation than inmates not in level IV? What numbers did you use to conclude this?

*Stata* can perform one other set of percentages, cell frequencies. Type:

```
.tabulate violation level4, cell
```

Again, you will see the counts, but now we have cell frequencies. All of the frequencies are out of the entire sample. For example, 247 inmates in the sample are in level IV security and also have committed violations. The cell relative frequency in this case will be 247/3922. (Remember there are 3922 inmates in our dataset). All cell relative frequencies are calculated by dividing the cell frequency by the total number of inmates in the dataset.

**Question 6:** Interpret three of the frequencies.

**Using one Categorical Variable and one Numerical Variable**

If we wanted to investigate if level IV inmates have higher classification scores than those not in level IV, we cannot use a two-way table since classification
score is not a categorical variable. Therefore we will use side-by-side box plots.

. sort level4
. graph score, box by(level4)

*Question 7: Describe the box plots. Do level IV inmates have higher classification scores (scores on the questionnaire) than those not in level IV security?*

Here is a list of Stata commands we used in this Analysis of Categorical Data lab. Use the space next to each command to make notes on what that command does.

*tabulate*
*tabulate, row*
*tabulate, col*
*tabulate, cell*
*sort*
*graph, box*
Assignment
Using the techniques and commands used throughout this lab, answer the following questions. Please give the commands used any plots used and how you reached your conclusions.

Are strike 2, or strike 3 inmates more likely to commit a violation? What percent of strike 3 prisoners are in level IV security? What percent of strike 2 prisoners are in level IV security? Do you feel the CDC’s placement policy with respect to managing the misconduct violations of level IV inmates, is effective?