**CPH 576C, group work**

**L2: Data cleaning and exploratory analysis**

0. Fill in the following table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type of data | Descriptive statistic | Graphical displays | Test comparing two independent groups | Test comparing two dependent groups | Test comparing multiple groups |
| Binary or Categorical | Frequency, Proportion | Bar charts (but not recommended) | Chi-square | McNemar’s (binary) | Chi-square |
| Continuous  (symmetric) | Mean, SD | Histogram | t-test | Paired t-test | ANOVA |
| Continuous  (skewed or not symmetric) | Median, IQR | histogram | Mann-Whitney (Wilcoxon Rank Sum) | Wilcoxon signed rank | Kruskall-Wallace |

1. Import the HELP data into the statistical software of your choice. Please sit next to someone who uses the same software so that you can help each other.

|  |  |
| --- | --- |
| **proc** **import** datafile = 'C:\CPH576C\data\HELP\help\_data.csv'  dbms = csv out = help replace;  **run**; | Can use either double or single quotes for path |
| import delimited using "C:\CPH576C\data\HELP\help\_data.csv" | Must use double quotes for path |

2. Get a quick description of the contents of the dataset

|  |  |
| --- | --- |
| **Proc contents data = help;**  **Run;** | To get variables in order, use the option varnum |
| codebook, compact | For more detail, leave off the compact option |

3. Consider the variables female, substance, cesd and i1. What type of variables are these, and what do they represent? Are there any values out of range?

|  |  |  |  |
| --- | --- | --- | --- |
| female | Binary | Sex: 0 = male, 1 = female | No values out of range (only 0’s and 1’s) |
| substance | Categorical | Drug of choice: alcohol, cocaine, heroin | No values out of range |
| cesd | Continuous | Depression | Possible range should be min = 0 and max = 20\*3 = 60. No vals out of range |
| I1 | Count | Average number of drinks per day | No values out of range (no negative values) but max seems very large |

4. What type of graph would you use to explore the association of

a. cesd and substance?

|  |  |
| --- | --- |
| graph box cesd, over(substance) | **proc** **sort** data = help; by substance; **run**;  **proc** **boxplot**;  plot cesd\*substance;  **run**; |
|  | |

b. cesd and i1?

|  |  |  |
| --- | --- | --- |
| graph twoway scatter i1 cesd | ods html;  **proc** **sgscatter** data = help;plot i1\*cesd;  **run**; | Proc gplot data = help;  Plot il\*cesd;  Run;quit; |
|  | | |