Problem Set 1
General Feedback

Nick Miller
17.871
3/4/2013
Comment your code!

• This is more than just /*Part 1*/, /*Part 2*/ etc. and more than /*Collapsing dataset*/
• Explain precisely what you are doing
• In future problem sets, I will take off points for lack of descriptive comments
Example

Good:
/*Generate a variable that equals 1 if the respondent reported waiting more than 30 minutes (when the variable= 4 or 5)*/

Bad:
/*Generating variable*/
Dealing with Missing Values

• Always use `tabulate` command (with the missing option) to get a sense of variables before analysis, and especially before transforming them or using them to create new variables.

• Missing values, or ‘don’t know’ answers on surveys should be treated as containing no information (i.e. should not be recoded as zero or as anything other than missing when transformed for use in other variables).
```
. tab q5

<table>
<thead>
<tr>
<th>mode of voting</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>in person on election day (at polling p)</td>
<td>6,103</td>
<td>64.93</td>
<td>64.93</td>
</tr>
<tr>
<td>in person before election day (early)</td>
<td>1,679</td>
<td>17.86</td>
<td>82.80</td>
</tr>
<tr>
<td>voted by mail (or absentee)</td>
<td>1,613</td>
<td>17.16</td>
<td>99.96</td>
</tr>
<tr>
<td>don't know</td>
<td>4</td>
<td>0.04</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>9,399</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

. tab q5, missing

<table>
<thead>
<tr>
<th>mode of voting</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>in person on election day (at polling p)</td>
<td>6,103</td>
<td>61.03</td>
<td>61.03</td>
</tr>
<tr>
<td>in person before election day (early)</td>
<td>1,679</td>
<td>16.79</td>
<td>77.82</td>
</tr>
<tr>
<td>voted by mail (or absentee)</td>
<td>1,613</td>
<td>16.13</td>
<td>93.95</td>
</tr>
<tr>
<td>don't know</td>
<td>4</td>
<td>0.04</td>
<td>93.99</td>
</tr>
<tr>
<td>.</td>
<td>601</td>
<td>6.01</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>10,000</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
```
Use caution when generating variables

• Stata treats the ‘.’ as higher than any integer value.
  
  • `gen newvar=1 if oldvar>5`
    – This will turn missing values on `oldvar` into 1s on `newvar`. This is bad.
  
  • `gen newvar=1 if oldvar>5 & oldvar!=.`
    – This will avoid that problem if missing values are dots (but not if they are 999, for example)
  
• Takeaway point: know how missing values are coded and tailor accordingly
More missing value pitfalls

• gen newvar1=0
  – This will code every observation as zero, regardless of missing data on existing variables. After all, Stata does not know in advance which old variable you will use to code the new variable

• gen newvar1=.  
  – This is generally safer because you start with a blank slate and recode based on specific values of an existing variable
Collapse command

collapse (mean) thirtyplus_2008 (count) n=
thirtyplus_2008 [aweight=weight], by(inputstate)

• The mean of a binary variable=the percentage of 1s. No additional variables/calculations needed.
• Counting the same variable used to calculate the mean ensures we are getting the correct sample size (need to generate a new variable in order to avoid duplicate names)