Valeria Salinas-Lopez

Fall 2018

Oct 9, 2018

SOC 3305

Lab 4

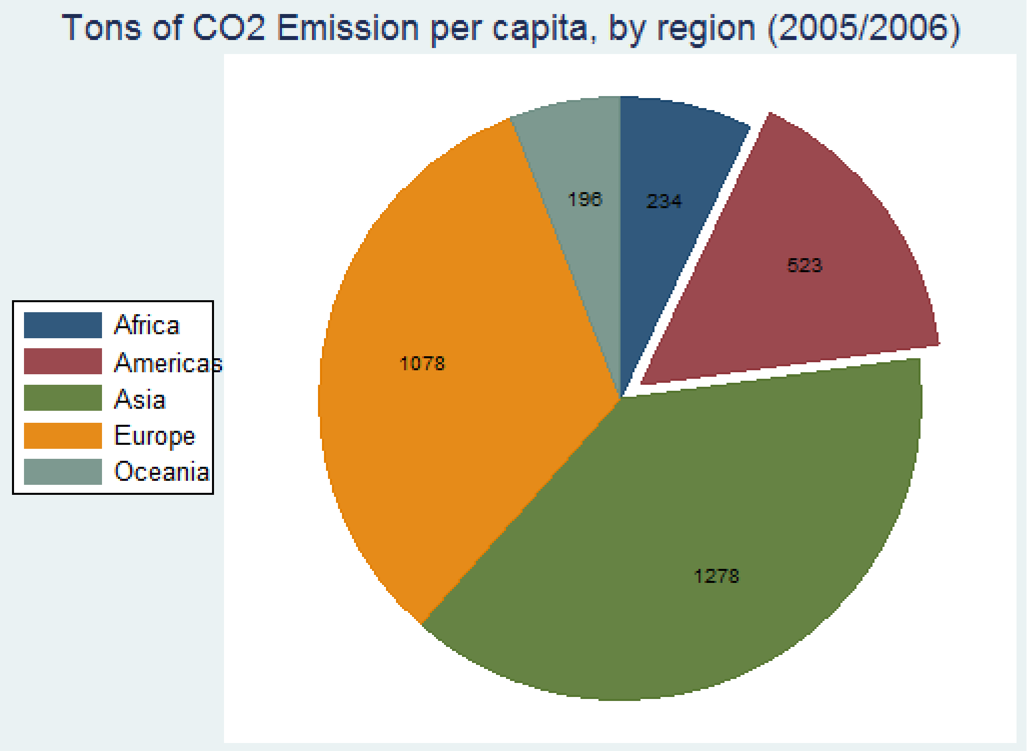
(Due in class on October 11, 2018)

Please save your final work as lastname_lab4.doc (put your name above the title page) and email it to instructor. Please explain your results with your own words. (You must use Stata program to do this lab). Prove your Stata command for each question.

1. **Please explain with your own words what a "pie chart" and "bar chart" are and how they are different from each other. Also give an example for each by using Nations2.dta. For pie charts use "co2" over region and explode "Americas". For bar charts use "co2" as well over region. Explain your charts. (20 pts).**

   Pie Chart: A pie chart is a circular graph that has pieces to a whole. Each piece represents the amount/proportion.

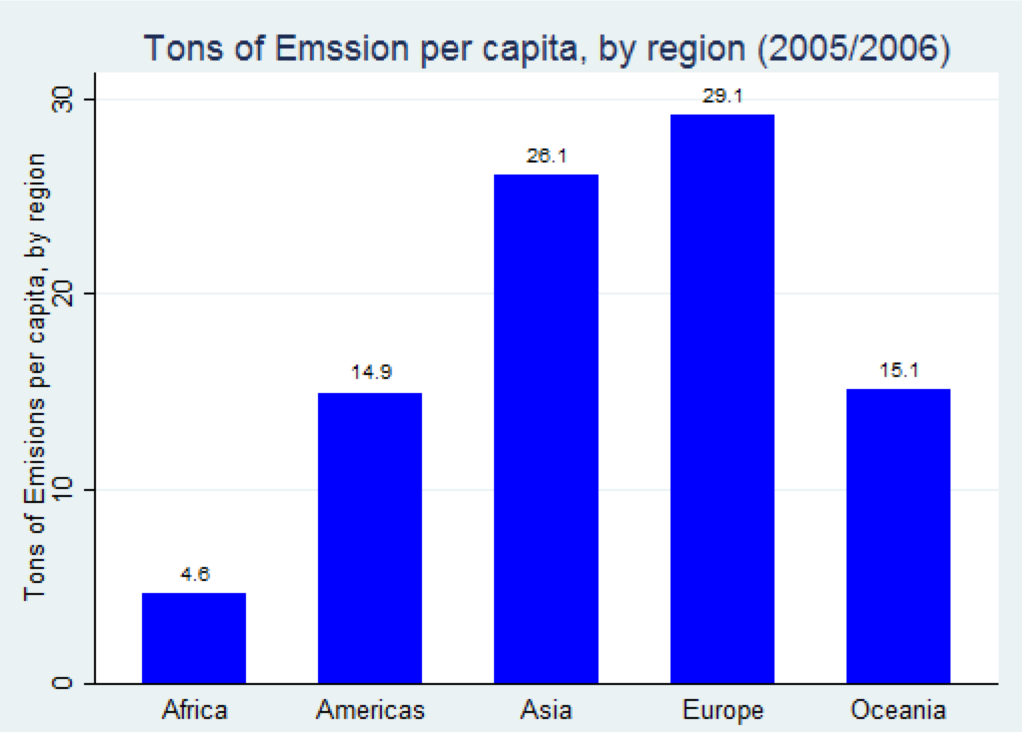
```
graph pie co2, over(region) pie(2, explode) plabel(_all sum, format(%4.0f)) title("Tons of CO2 Emission per capita, by region (2005/2006)") legend(col(1) position(9))
```



Explanation: The Americas is the 3rd largest emitter of CO2 in tons. The largest emitter of CO2 is Asia followed by Europe. Oceania and Africa are close to the amount of CO2 that they emit.

Bar Chart/Graph: A bar chart/graph is a way to show categorical data. There are rectangles whose length represents the proportional value of that specific category.

```
graph bar co2, over(region) ytitle("Tons of CO2 Emission per capita, by region (2005/2006)")
blabel(bar, format(%3.1f)) bar(1, color(blue)) bar(2, color(orange)) legend(ring(0)
position(11) col(2) label(1 "Mean") label(2 "Median") symxsize(*.5))
```



Explanation: Europe and Asia are still the top emitters. Americas and Oceania are close together and Africa is the least.

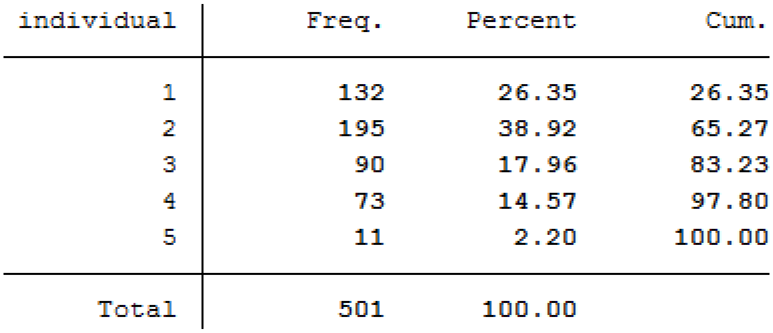
Differences: Pie charts can only compare the different parts of the same whole, while the bar charts/graphs can compare different things since they show different categories.

2. **Explain svyset and also weighting in survey analysis. What is the difference between "pweight" and "fweight"? Also explain difference between "n" and "N" in survey analysis. (20 pts)**

   Svyset is used to declare survey design for the dataset. **Weighting** is used in order to have more accurate data. If we have more women (56% of the pop.) and the less men (44% of the pop.), then we need to make sure the observations are weighted based on the percent of each sex. **"Pweight"** allows for Stata to use sampling weight as the number of subjects in the population that each observation represents. **"Fweight"** are used to indicate the number of times the observations actually happened. **"N"** is the population size while **"n"** is the sample size.

3. **Use Granite2011_6 data from class dataset. Generate new variable "individual" in which it includes both adults and children in each household. Design a new weight for your new "individual" measurement according to each person including adults. See how results change representation number of individuals in each household. Provide your tab before and after weight and interpret your results. (20)**

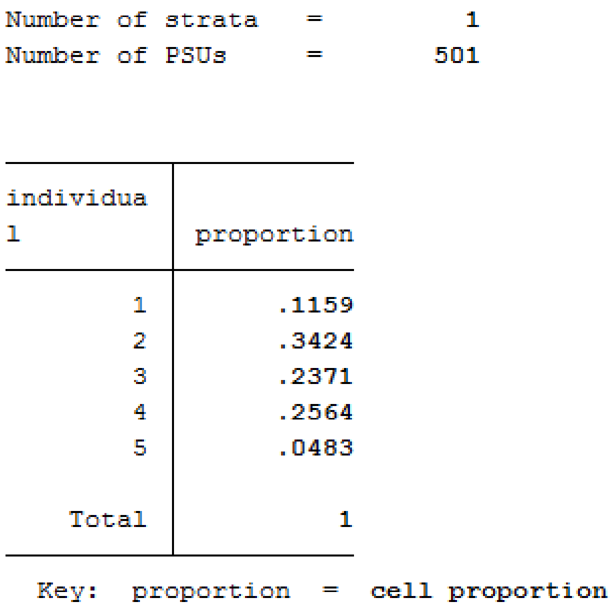
   Command: I used generate individual = adults+ children to create the new variable of "individual." Then I used tab individual to tabulate the number of adults and children.



```
 individual |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        132       26.35       26.35
          2 |        195       38.92       65.27
          3 |         90       17.96       83.23
          4 |         73       14.57       97.80
          5 |         11        2.20      100.00
------------+-----------------------------------
      Total |        501      100.00
```

In order to check if the data was accurate, I calculated the percentage of the amount of individuals that have one person in the household and compared it to the percentage on tab 1 (above). When I do this 132 (total of one household individuals in tab 1) divided by 1139 (the actual total of individuals) I get 0.116. This means that the true percentage for the individuals with one person in a household should be 11.6%, not the 26.4% that is shown in tab 1.

Next I used, generate individualwt= individual*(501/1139), and svyset _n[pw=individualwt] and then finally svy: tab individual, percent to compare the true data to the original data.



```
Number of strata   =         1                 Number of obs     =        501
Number of PSUs     =       501                 Population size   = 501.000007
                                               Design df         =        500

----------------------
individua |
1         |  proportion
----------+-----------
        1 |      .1159
        2 |      .3424
        3 |      .2371
        4 |      .2564
        5 |      .0483
          |
    Total |          1
----------------------
  Key:  proportion  =  cell proportion
```

Based on the second tab, you can see that the one-person households were originally overrepresented compared to the true data. Two-person households were overrepresented as well. The three-person, four-person, and five-person households were all underrepresented.

4. **Use Granit2011_6.dta and assume that 2010 census provided that females compose 48% of NH population while male 52%. How would you incorporate to your post-stratification weights? (20)**

   In the 2010 Census, it states that 45.35% of males ("0") responded to this survey. It also states that 54.65% of women responded. Since women make up 48% of the population and men make up 52% of the population, men are underrepresented and we need to apply post-stratification weights in order to get more accurate data. The following commands would help do that.

   Commands: generate sexwt= 52.0/45.35 if sex==0

   generate sexwt= 48.0/54.65 if sex==1

   After that, you have to apply these weights to your data. I used svyset [pw=sexwt].

5. **Explain what logistic regression is and when to use. Run a regression predicting probability of a vote for "Obama" in 2008 election by whether or not graduated from college , gender (sex) and belief in shrinking ice on the Arctic Ocean's surface. Explain significance of your variables and relation between your independent and dependent variable. (20 pts)**

   A logistic regression helps to predict the probability of a dependent variable based on one or more independent variables. You would use this whenever you were trying to see the probability of something happening based on one variable you have.

Command: logit obama college sex warmice

```
Iteration 0:   log likelihood = -354.15026
Iteration 1:   log likelihood = -332.72628
Iteration 2:   log likelihood = -332.67745
Iteration 3:   log likelihood = -332.67745

Logistic regression                               Number of obs     =
>        511
                                                  LR chi2(3)        =
>      42.95
                                                  Prob > chi2       =
>     0.0000
Log likelihood = -332.67745                       Pseudo R2         =
>     0.0606

─────────────┬──────────────────────────────────────────────────────
> ──────────
       obama │      Coef.   Std. Err.      z    P>|z|     [95% Conf.
> Interval]
─────────────┼──────────────────────────────────────────────────────
> ──────────
     college │   .7685804   .1858381     4.14   0.000     .4043445
>   1.132816
         sex │   .4276524   .1857039     2.30   0.021     .0636793
>   .7916254
     warmice │  -.3778821   .0955251    -3.96   0.000    -.5651078
>  -.1906564
       _cons │  -.0901496   .2285452    -0.39   0.693    -.5380899
>   .3577907
─────────────┴──────────────────────────────────────────────────────
> ──────────
```

Obama and college: College is significant variable because P is less than 0.05. Voting for Obama is not highly influenced by graduating from college. I know this because the coefficient is a very low number. Therefore, Obama votes are not dependent on college graduation. They are positively correlated, which means that as college graduation increases so does Obama.

Obama and Sex: Sex is a significant variable because P is less than 0.05. Voting for Obama is not highly influenced by sex since the coefficient is small. Therefore, Obama votes are not dependent on sex. The two variables are positively correlated, which means as one goes up so does the other.

Obama and Warmice: Belief in shrinking ice is significant because P is less than 0.05. Voting for Obama is not highly influenced by the belief in shrinking ice because the coefficient is small. Therefore, Obama votes are not dependent on belief in shrinking ice. These variables are negatively correlated, which means as one goes up the other goes down. As the belief in shrinking ice goes up, the vote for Obama goes down, and vice versa.