Madeleine Morlino
INTLPOL 301A

Assignment #1

Part A

1) Descriptive Statistics Table

Variable	Obs	Mean	Std. Dev.	Min	Max
year	2,867	2016	0	2016	2016
health	2,867	1.410882	1.237355	0	9
region	2,867	5.089292	2.481976	1	9
rincome	2,867	6.798047	8.509837	0	98
income	2,867	15.21904	18.47239	1	98
race	2,867	1.364144	.6518624	1	3
sex	2,867	1.554935	.4970596	1	2
educ	2,867	14.00453	5.608504	0	99
age	2,867	49.32961	17.90478	18	99
marital	2,867	2.666899	1.685587	1	9
occ10	2,867	4171.778	2683.359	0	9999
hrs1	2,867	23.86467	24.28167	-1	99
wrkstat	2,867	3.036972	2.333057	1	9
id_	2,867	1434	827.7759	1	2867
uscitzn	2,867	.2322986	.827664	0	9

2) A. There are 328 immigrants in the sample. The people who answered yes to the following categories were counted as immigrants: U.S. citizen and not a U.S. citizen. Specifically, people who answered that they were a U.S. citizen were naturalized immigrants, and people that said they were not a U.S. citizen were immigrants. In accordance with the GSS Codebook, non-applicable corresponds to U.S. citizens, since these individuals answered that they were born in the U.S. in the previous survey question. Being born in the U.S. automatically means that these individuals were U.S. citizens.

   B. The following variables were created: immigrant, American born, and unknown data. American born individuals included categories 0,3, and 4: not applicable, a U.S. citizen born in Puerto Rico, the U.S. Virgin Islands, or the Northern Marianas Islands, and born outside of the United States to parents who were U.S. citizens at that time, respectively. These specific categories were selected for this category, since American born implies automatic citizenship upon birth; all the aforementioned categories provide those individuals automatic U.S. citizenship. Unknown data include categories 8 and 9: don't know and no data, respectively.

	Under of the Age of 30	
	Immigrant	American Born
Percentage	14.94%	17.00%

The data shows that there are more American born individuals under the age of 30 in comparison to the same age group for the immigrant population.

C.

	Average Education Attainment (years)	
	Immigrant	American Born
Mean	13.28	13.81
Median	14	13

The mean values show that the educational attainment was only slightly higher for American born citizens in comparison to the educational attainment for immigrant population. Looking at the median values, the immigrant population median value was larger than the mean value; this means that for immigrants, the data was right skewed. Conversely, for the American born population, the median value was smaller than the mean value, which means the data was left skewed.

D.

	Average Number of Hours Worked Last Week (hours)	
	Immigrant	American Born
Mean	40.92	40.89

As seen by the table above, the average number of hours worked last week for both immigrants and American born individuals are similar at 40.92 and 40.89 hours, respectively.

** Please reference the do file with the specific commands and notes for all of Part A*

Part B

Question 1:

A. The resulting sample is not a simple random sample, since every possible student did not have an equal chance of being selected. Specifically, the implemented quota meant that all the undergraduate students at the University of Florida did not have the same chance of inclusion in the sample.
B. The sample selected is a stratified sample, since the student population was randomly selected and separated into strata based on race. The key advantage of this type of sample over a simple random sample is that the sample has an equal representation of both Black and White racial groups. Other advantages include better data precision and accuracy, and minimizing biases when selecting the samples.

C. It would be advantageous to incorporate a clustering component to the sampling design, because clustered random samples offer the advantage of easier logistics and cost. Since the University of Florida undergraduate population size is large, clustering allows for economical and practical benefits. Clustering by residential halls/dormitories would be an ideal criteria, because this geographic boundary allows for adequate representation of the entire population under observation.

Question 2:

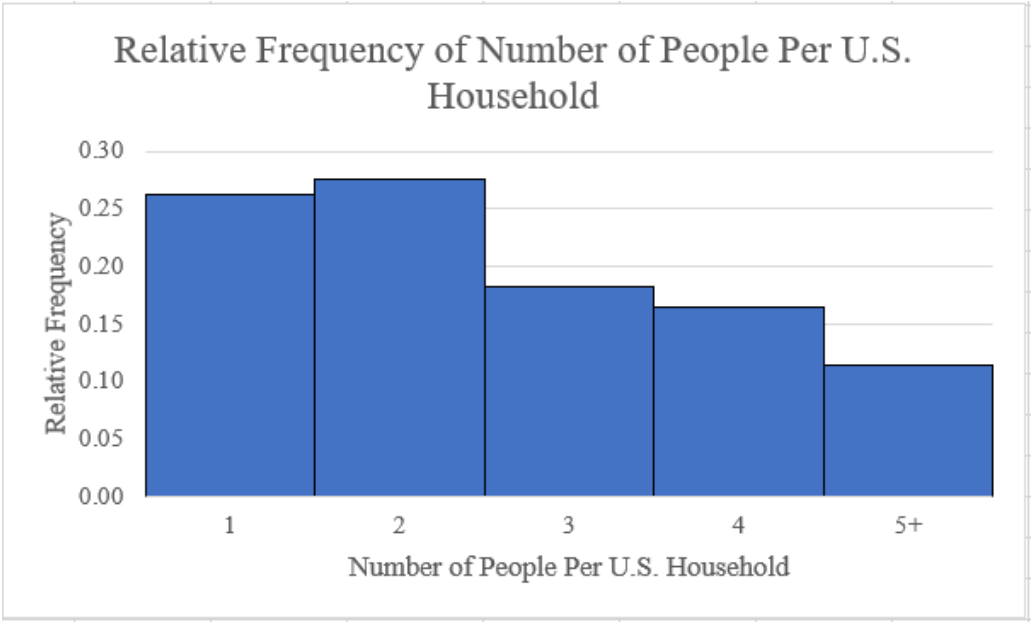
A stratified random sample is the most advantageous probability sampling design, since dividing the population into strata based on race would guard against an unrepresentative sample. Specifically, this type of sampling design ensures that all racial groups are accounted for, thereby creating more precise and accurate data.

Question 3:

A.

Relative Frequency Distribution		
Number of People Per Household	**Frequency (millions)**	**Relative Frequency**
1	28.5	0.26
2	30.0	0.28
3	19.8	0.18
4	17.9	0.16
5+	12.4	0.11

B.



The histogram's shape is right skewed as seen by the histogram's long right tail.

C. The median household size is two, meaning that two people per household was the observation falling in the middle of the ordered sample. The mode household size is two, since two people per household was the value that occurred most frequently; this is visually demonstrated by the histogram above.

Question 4:

Table 1: Life Expectancy for Females and Males in All Countries

	Female	Male
Mean	74.67	69.83
Variance	60.76	53.86
Standard Deviation	7.79	7.34
Coefficient of Variation	0.10	0.11

Table 2: Life Expectancy for Females and Males in Countries Currently Belonging to the European Union

	Female	Male
Mean	82.96	77.20
Variance	3.75	11.97
Standard Deviation	1.94	3.46
Coefficient of Variation	0.02	0.05

Table 3: Life Expectancy for Females and Males in African Countries

	Female	Male
Mean	65.91	61.94
Variance	39.22	32.94
Standard Deviation	6.26	5.74
Coefficient of Variation	0.10	0.09

** Somalia was not included in the calculations, since there was no complete data provided for female and male life expectancies*

As seen in the Table 1 above, all countries in the world have significantly larger variance and standard deviation values in comparison to countries belonging to the European Union and African countries. With variance and standard deviation describing the spread of the data, this means that the range of values for life expectancy for males and females

in all countries is more spread out in comparison to Tables 2 and 3. The large standard deviation value and coefficient of variation also indicate more variability of the mean. This can be attributed to the presence of more extreme values, which will influence the coefficient of variation.

As seen in Table 2, countries belonging to the European Union have the highest life expectancy mean for both genders and the lowest standard deviation, variance, and coefficient of variation. This means that there is less variability of life expectancy for countries belonging to the European Union. This could be attributed to the presence of more developed countries in the region, resulting into better access to technology, quality healthcare and medical services.

As seen in Table 3, females and males in African countries have the lowest mean value of life expectancy when compared to Tables 1 and 2. This means that individuals in this part of the world are expected to live the shortest. This could be attributed to the presence of more developing countries in the region, resulting into limited access to technology, quality healthcare and medical services. There is more variability for life expectancy within African countries in comparison to countries belonging to the European Union.

Consistently across the data, females have a longer average life expectancy in comparison to their male counterparts.

Question 5:

A. The average stay for people in the new study is less in comparison to the average stay for people in the old study. This is seen by the new study's mean value of 17 days, which is lower than the old study's mean of 28.1 days. The median value for the new and old studies are 12 and 23.5 days, respectively. With this value being insensitive to the distances of the observations from the middle, this measure of central tendency similarly indicates that people are now spending less time in the facility. This could be attributed to factors such as the development of more effective methods for treating alcoholism in addition to technological and scientific advancements. Since both the new and old studies have higher means than medians, the studies are positively skewed.

There is slightly more range and variability in how long patients stay in the facility in the new study when compared to how long patients stay in the facility in the old study. This is seen when comparing the range and standard deviations of the two studies. The range for the new study was 38 days compared to 37 days for the old study; the new study also has slightly higher standard deviation of 13.23 days in comparison to the old study's standard deviation of 12.68 days.

B. It is possible to calculate the median; since the patient's length of stay is a minimum of forty days, this value represents a value close to the higher end of the new study range.

The median is not influenced by the higher end of the range, and instead relies on the relative position of the values. It is not possible to calculate the mean, since the mean relies on knowing the exact values as opposed to the values relative position in order to be calculated.

Question 6:

A. The data suggests a right skewed distribution, because the mean values are greater than the median values.

B. The overall mean income is $43,803.11

	Median	Mean	Number of ppl	
Females	23407	33915	112,000,000	
Males	37374	53871	110,000,000	
			222,000,000	
Overall Mean Income	**43803.11**			
Calculation	((112,000*33915)+(110,000,000*53871)/222000000			

Question 7:

The statistics suggest that income inequality increased between 1989 and 2013, since there was more variability in the distribution of family net worth. Since the mean value drastically increased but the median value slightly decreased, the distribution became even more skewed to the right. The higher mean indicates that wealthier families became wealthier; the decrease in the median indicates that half the U.S. population's family net worth was less in 2013 than in 1989.

Question 8:

A. Y is a discrete variable, because the number of languages in which a person is fluent is quantitative data that takes on countable, whole number values.

B.

Probability Distribution of Number of Languages in Which A Person Is Fluent			
Fluency in Number of Languages (X)	0	1	2
Probability P(X)	0.01	0.83	0.16

C. The probability that a Canadian is not multilingual is 0.84

Probability Distribution of Number of Languages in Which A Person Is Fluent			
Fluency in Number of Languages (X)	0	1	2
Probability P(X)	0.01	0.83	0.16
	0	0.83	0.32
Probability that a person is not multilingual		84%	
		^SUM(B24:C24)	

D. The mean of this probability distribution is 1.15; this means that the average Canadian only speaks one language and is therefore, not multilingual.

Probability Distribution of Number of Languages in Which A Person Is Fluent				
Fluency in Number of Languages (X)	0	1	2	
Probability P(X)	0.01	0.83	0.16	**MEAN**
	0	0.83	0.32	1.15
				^SUM(B25:D25)

Question 9:

A. The mean SAT for the entire population of test-takers in 1981 was 507.59; the mean SAT for the entire population of test-takers in 2002 was 507.93.

	1981			2002		
	SAT score	n	% of Test-Taking Population	SAT score	n	% of Test-Taking Population
White	518	734,395	84.80%	525	676,141	68.98%
Black	432	75,651	8.74%	440	128,698	13.13%
Asian	490	29,213	3.37%	514	103,137	10.52%
Latino	454	22,069	2.55%	459	65,076	6.64%
Native American	459	4,721	0.55%	474	7,213	0.74%
		866,049.00			980,265.00	
MEAN	507.590768		**MEAN**	507.9263352		
Calculation	(439,598,477)/866,049		Calculation	(497,902,409)/980,265		

B. The change in SAT scores between 1981 and 2002 for each racial/ethnic group does not match with the 1981-2002 change for the entire population of test takers, because of the significant changes in how many students in each racial group made up the entire test-taking population. Specifically, White students made up 84.8% of the data selected in the 1981 study; this high percentage of White students significantly brought up the average SAT score. However, in 2002, despite the other ethnic groups (Black, Asian, Latino, and Native American) increasing their score and number of people taking the test, the percentage of White students dropped to 68.98%. This meant that the White student's higher SAT score average ended up balancing with the other ethnic groups' scores.

	1981			2002		
	SAT score	n	% of Test-Taking Population	SAT score	n	% of Test-Taking Population
White	518	734,395	84.80%	525	676,141	68.98%
Black	432	75,651	8.74%	440	128,698	13.13%
Asian	490	29,213	3.37%	514	103,137	10.52%
Latino	454	22,069	2.55%	459	65,076	6.64%
Native American	459	4,721	0.55%	474	7,213	0.74%