Logan Albiani

2/18/18

Prof. Jones

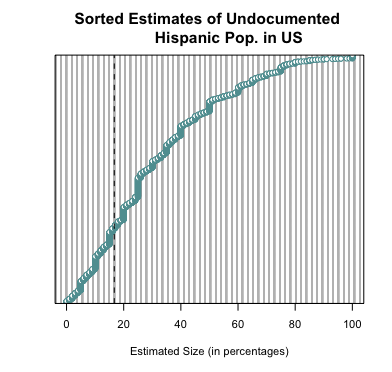
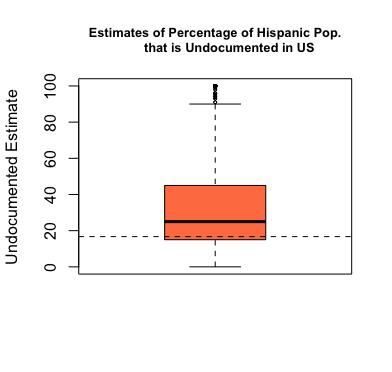
POL 051

*Homework 1:*

**Problem 1:**

As can be seen below, I’ve created both a boxplot (right) which shows the estimates of the Hispanic population which is undocumented. We can see a number of various pieces of information. We have the 25th percentile, represented by the bottom of the orange box, is an estimate of 15% and the 75th percentile, represented by the top of the orange box, is given at 45%. We can also see the median which is shown by the solid black line in the box, which in this case is 25%. We’re also given the interquartile region, which is the range in which 50% of the data lie, by the orange box itself. We can also derive the minimum and maximum values of our data set which are given by the end of each whisker as well as outliers, which are the small circles outside of the chart itself. The estimate for the true population of the Hispanic Population which is undocumented has also been included in the example as dotted line which goes through the orange box.

Next, we’ve the dotchart which is in itself simply all of the responses given in the data set



sorted from the lowest estimates to the highest estimates. As one can clearly gather from the chart, the large majority of those estimates were overestimates of where the true population of 16.7%. However, with this being said, one can also see half of the data, like the boxplot, is between 15% and 45%.

Still, it is worth repeating that both of these charts show generally incorrect estimates by the participants. Though the true population lies around 16.7%, around 73% of the participants overestimated the size of the undocumented population

**Problem 2:**

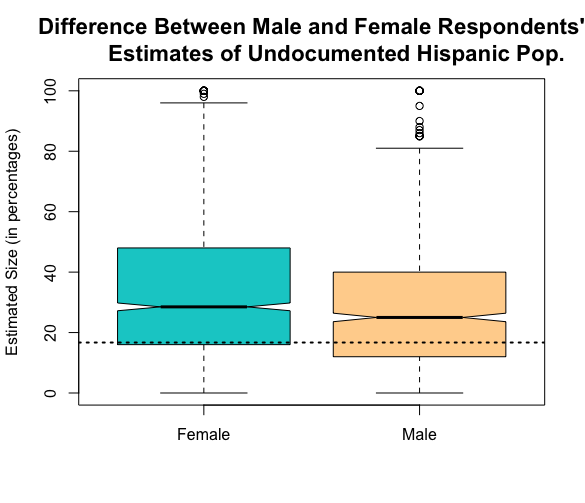
A.) The null hypothesis will be that there is no difference between the estimates of males and females whereas the alternative will be that the estimates of males and females do not equal one another. I’ve chosen these two hypotheses because the null hypothesis is chosen in a manner that assumes there to be no difference between outcomes based on some variable (gender in this case) and the alternative was chosen in a manner so to allow there to be the use of a two tailed hypothesis test since I don’t know whether or not one genders’ estimates will be generally larger or smaller than the others’.

B.) Table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gender** | **Mean** | **Median** | **Standard Deviation** | **Inter-Quartile Range** |
| **Male** | 29.00 | 25.00 | 22.02 | 28.00 |
| **Female** | 33.71 | 28.50 | 22.21 | 32.00 |

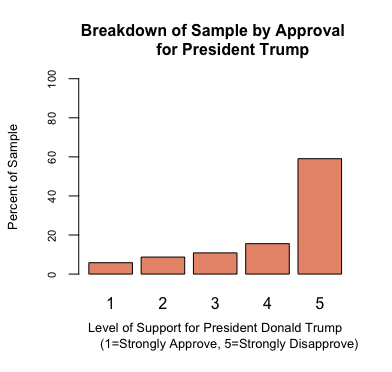
C.) As can be seen below, I’ve created two boxplots comparing male responses and female responses. We can see in these two charts that both the male and female participants overestimated the size of the undocumented Hispanic population; however, we can see from this data set, that female participants had a higher estimate of the undocumented size than did male participants. This is the case because the females in this study had a higher mean, median, and range of interquartile values. This conclusion that males and females do not in fact estimate the population of undocumented Hispanics in the United States is consistent with my alternative hypothesis that males and females do not estimate undocumented populations to be the same.

D.) Using a 95% confidence interval, a T-test comes back to conclude that the p-value of for the populations being unequal to one another by random chance is .0000001624, which is short to say, it is very unlikely. This p-value gives us evidence to support the alternative hypothesis that males and females estimate populations of undocumented Hispanics differently and more specifically, females overestimate the undocumented Hispanic population. This is due to the fact that the p-value given is such an astronomically small number that is hard to find evidence against this result happening purely by coincidence rather than by some type of correlation.



**Problem 3:**

A.) The plot below shows breakdown of approval for Donald Trump by percentages the sample. On the x-axis, I have the levels of support for President Trump denoted 1 – 5 (1 being strongly approve, 5 being strongly disapprove). As can be gathered by looking at the bar graph, most respondents indicated strong disapproval for the President while those respondents indicating strong or even moderate approval of President Trump composed a small minority (approximately 17%) of the sample.



**Problem 4:**

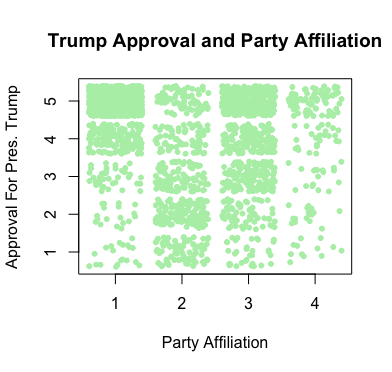
A.) Given that we expect there to be a partisan attachment to the approval of Donald Trump, and we’d expect those who self-identify as Democrats to more readily disapprove of the President, whereas Republicans would be more hesitant. The null hypothesis in this case then would be that there are no differences between the approval and disapproval ratings due to a partisan difference between respondents. Our alternative hypothesis is that Democrats and

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Party ID** | **Mean** | **Median** | **IQR** | **Standard Deviation** |
| **Democrat** | 4.77 | 5.00 | 0 | 0.72 |
| **Republican** | 2.812 | 3.00 | 2 | 1.35 |
| **Independent** | 3.96 | 4.00 | 2 | 1.19 |
| **Other** | 3.93 | 5.00 | 2 | 1.35 |

Republicans in the sample will not have the same levels of approval of Donald Trump.

B.) As we can see in the table below, the table reveals that indeed, the data are consistent with the alternative hypothesis. Democrats have a much higher disapproval of Trump than do Republicans. In fact, in both the mean and the median, are far different. Republicans indicate that they have neither an approval or disapproval of the President while Democrats indicate, in both the mean and median, that they have a strong disapproval of Donald Trump. This is consistent with our alternative hypothesis because it stated that there would be a partisan difference between the Democrats and Republicans on the favorability of Donald Trump. Interestingly, Partisan identifiers look to be right in between the Democrats and the Republicans. Their mean and median are approximately a moderate disapproval, which does indeed make sense when we think about the independents on a left right spectrum. These people fall, ideologically, between Democrats and Republicans, so it only makes sense that their mean and median approvals of Trump would fall directly between Republicans and Democrats.

C.) Below, we’ve a jitterplot. In this plot, we can see, along the y-axis, the approval ratings of Donald Trump (5 being strong disapproval, while 1 being strong approval) and on the x-axis, we’ve party affiliation (1 being Democrat, 2 being Republican, 3 being Independent, and 4 being some other party). As we can see in this plot, each dot represents a respondent, and their location is based off of their party ID and their approval ratings. We see a strong concentration of 4’s and 5’s amongst Democrats while 1’s, 2’s, and 3’s are sparsely populated. In the Republican category, we see less density in each category, however, those blocks which are most concentrated are 1’s and 2’s. This is a reflection of the previously stated data, which was that Democrats have a stronger disapproval of Donald Trump where as Republicans generally neither approve or disapprove of the President’s performance. Independents and other party identifiers tend to have a mix between Republican and Democratic levels of approval, which is also a reflection of the previously given data.



D.) The results of the two-sided t-test are that the p-value of the means between the party identifiers being different by random chance, with 95% confidence, is .99, which is greater than what we would need to accept the null hypothesis (.95). Therefore, we reject the null hypothesis because the p-value indicated is statistically significant.

**Problem 5:**

A.) In this example, we have created a random number of samples from the given data and allowed for random sampling amongst these different sets of observations (10, 50, 500, and 1000 different sets of observations). As can be seen by the estimates in the table below, as the number of random samples increases, the closer the mean approaches the true population mean of 31.9. The standard deviation also steadily increases as the number of observations increases. This shows us that when the data pool is increased in size, we have a more reliable data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **# of Observations** | **10** | **50** | **500** | **1000** |
| **Mean** | 27.89 | 28.50 | 28.92 | 31.51 |
| **Standard Deviation** | 17.60 | 18.30 | 19.86 | 21.64 |