Christopher Messina

March 13, 2019

Assignment #4

*Step 1:* *Downloaded*.

*Step 2: Classifying Protected Class Variables*

Protected Class Classifications:

* Age: 'old','older','young','younger','teenage','millenial','middle aged','elderly'
* Gender: 'transgender','trans','queer','nonbinary','male','female'
* Religion: 'christian','muslim','jewish','buddhist','catholic','protestant','sikh','taoist'
* Nation of Origin: 'african', 'european', 'mexican', 'canadian', 'american', 'asian', 'indian', 'middle eastern', 'chinese', 'japanese'
* Disability: 'blind','deaf','paralyzed'
* Race: 'black', 'african american','hispanic', 'white','latino','latina','latinx'

*Step 3: Correlation Coefficients*

|  |  |  |
| --- | --- | --- |
| Protected Class | Correlation Coefficient | Strength of Correlation |
| Age | 0.0004815827683607928 | Very Weak |
| Gender | 0.0025885021950986294 | Very Weak |
| Religion | 0.000585583912977551 | Very Weak |
| Nation of Origin | -0.00044578509613503865 | Very Weak |
| Disability Status | 0.006234738068585817 | Very Weak |
| Race | 0.0021587300762926187 | Very Weak |

*Step 4: Calculate Population Mean and Standard Deviation*

Population Mean: 0.5514054399634942

Population Standard Deviation: 0.361704437963752

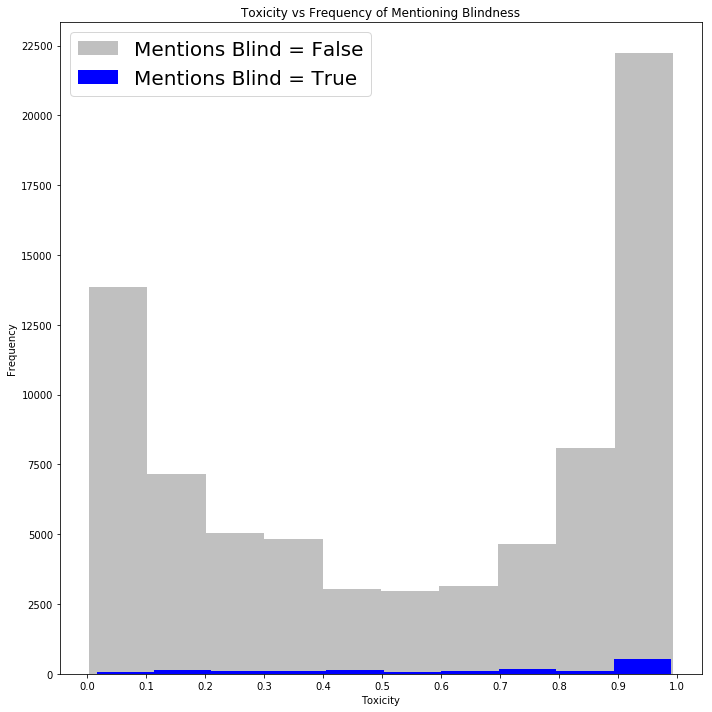
The range of values around the mean that includes 95% of toxicity (i.e. two standard deviations away from the mean in each direction) is (-0.1720034359640098, 1.2748143158909984).

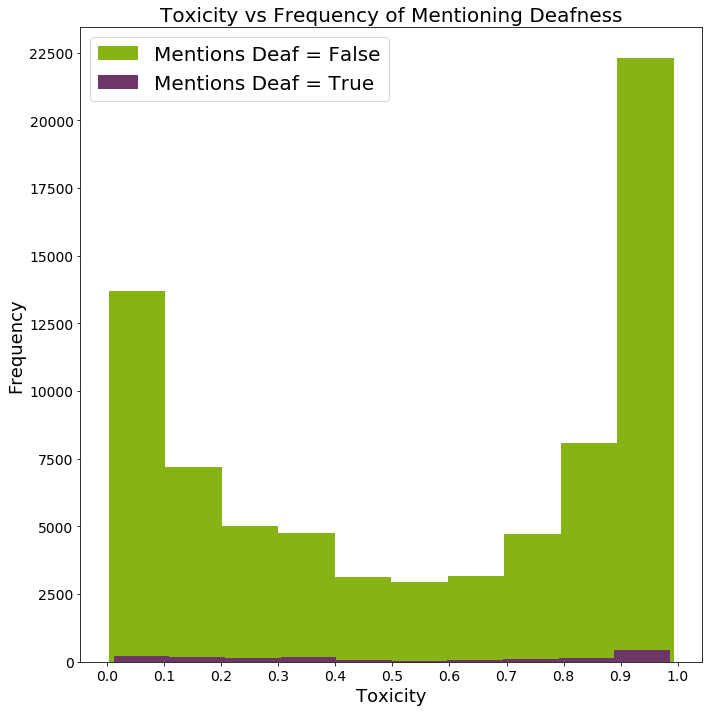
|  |  |  |  |
| --- | --- | --- | --- |
| Percentage of Pop. | Mean | Standard Deviation | Margin of Error (1/ ) |
| 10% | 0.5369160240630795 | 0.3628197959335086 | 0.011428011702929945 |
| 25% | 0.543186444501071 | 0.36353787764264306 | 0.007227992418765111 |
| 50% | 0.5457585704758895 | 0.36228545933812784 | 0.005110962453673766 |

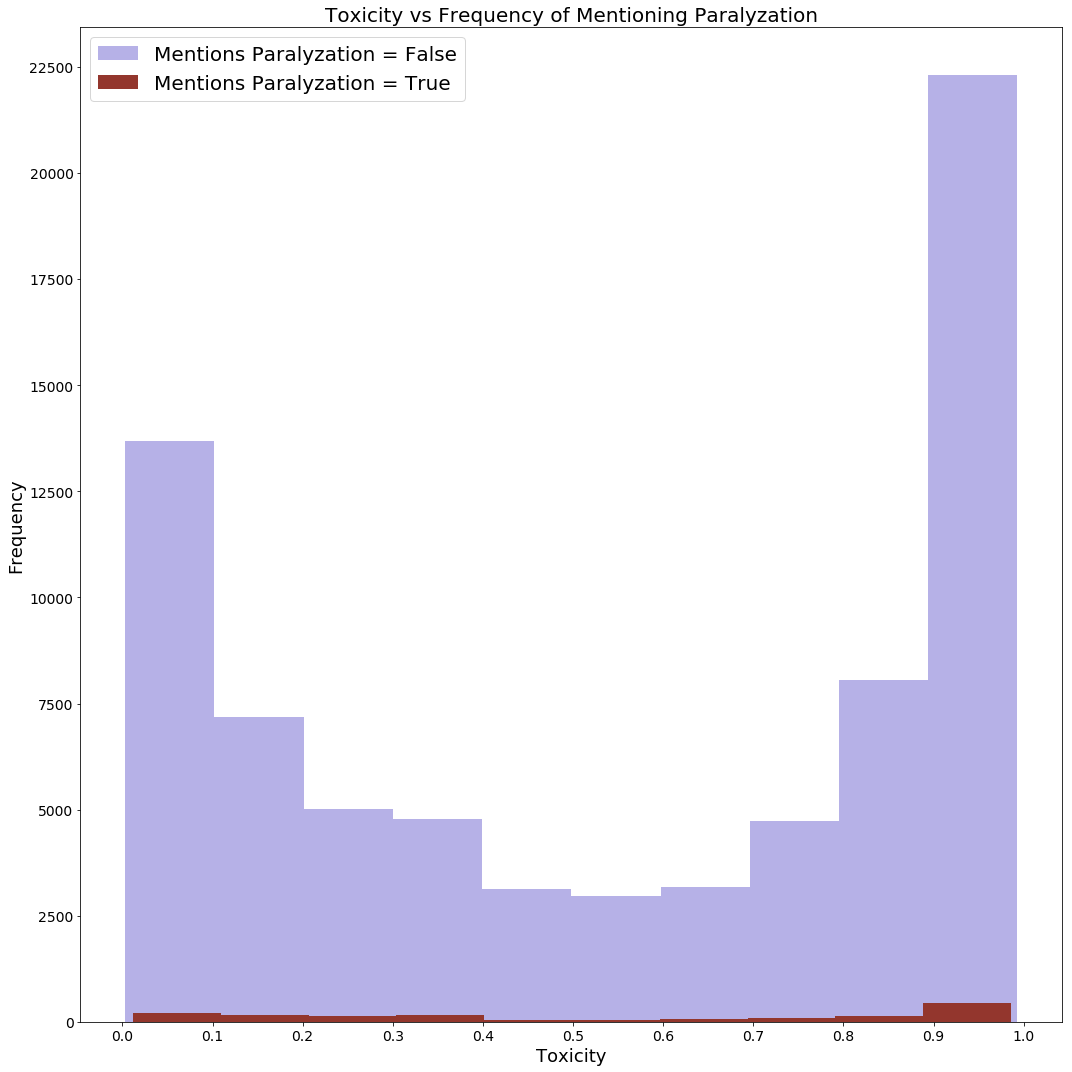
*Step 5: Graphing Toxicity and Frequency for Each Subgroup*

Protected Class chosen – Disability Status

Subgroups – Blind, Deaf, Paralyzed (histograms presented in this order)







*Step 6: Random Sampling Toxicity*

* Calculate the mean and standard deviation of TOXICITY associated with the protected class group (Hint: TOXICITY values should only be included in the calculation when the associated protected class group value is TRUE). Run the random sampling method using 25% and 50% of the data. For each, what is the mean and standard deviation? Indicate (yes/no) if the values lie within the associated population margin of error.
* Calculate the mean and standard deviation of the TOXICITY of the protect class group (Disability status)

mean: 0.5824167182877587

standard dev: 0.33493112044955553

margin of error: 0.014838036499774766

* Mean +- margin of error leads to a range of (0.56757868178, 0.59725475478) for acceptable sample means
* Calculate the mean and standard deviation of 25% and 50% using the random sampling method
  + 25%:

mean: 0.5774309063767605

standard dev: 0.3418490070040604

* + 50%:

mean: 0.5727746072351386

standard dev: 0.3400257656735319

* Both of these means fall within the associated population margin of error range that I have defined above.

*Step 7: Random Sampling Subgroup Toxicity*

* Calculate the mean and standard deviation of TOXICITY associated with each subgroup that is a member of the protected class group (Hint: TOXICITY values should only be included in the calculation when the associated subgroup value is TRUE). Run the random sampling method using 25% and 50% of the data. For each subgroup, what is the mean and standard deviation? Indicate (yes/no) if the values lie within the associated population margin of error.
  + Blind:
    - Population:

mean: 0.6369212232978864

standard dev: 0.3079153691402614

margin of error: 0.02570023310217136

lower limit: 0.611220990195715

upper limit: 0.6626214564000578

* + - 25%:

mean: 0.6416877455699208

standard dev: 0.31490206340210664

* + - 50%:

mean: 0.6116997958731836

standard dev: 0.314087076173162

* + - Yes, both samples fall within the acceptable margin of error range defined above.
  + Deaf:
    - Population:

mean: 0.5551644657826948

standard dev: 0.3444332883044277

margin of error: 0.02570023310217136

lower limit: 0.5294642326805235

upper limit: 0.5808646988848662

* + - 25%:

mean: 0.559847816717678

standard dev: 0.3501721681874665

* + - 50%:

mean: 0.5436430317582563

standard dev: 0.34181126908345616

* + - Yes, both the 25% and 50% samples fall within the acceptable margin of error range defined above.
  + Paralyzed:
    - Population:

mean: 0.5551644657826947

standard dev: 0.34443328830442765

margin of error: 0.02570023310217136

lower limit: 0.5294642326805233

upper limit: 0.5808646988848661

* + - 25%:

mean: 0.559847816717678

standard dev: 0.3501721681874665

* + - 50%:

mean: 0.5436430317582563

standard dev: 0.34181126908345616

* + - Yes, both the 25% and 50% samples fall within the acceptable margin of error range defined above.

*Step 8: Stratified Random Sampling Subgroup Toxicity*

* Run the stratified random sampling methods that covers 50% of the data. What is the mean and standard deviation? For each subgroup, what is the mean and standard deviation? Indicate (yes/no) if the values lie within the associated population margin of error.
* Method: Selected a random 50% of the data that fell within the protect class group, Disability Status. I found the mean and standard deviation of that 50% of the data. From there, I calculated the mean and standard deviation of each subgroup that fell within this 50% data population.
  + Population:

mean: 0.5679567756085425

standard dev: 0.336948390088002

margin of error: 0.020984152456968483

lower limit: 0.546972623151574

upper limit: 0.588940928065511

* + Blind:

mean: 0.6191523962736841

standard dev: 0.3104120852574086

* + Deaf:

mean: 0.5417550876698113

standard dev: 0.34811730510474953

* + Paralyzed:

mean: 0.5426420561612484

standard dev: 0.34530472462397604

* + The mean for the Blind sample falls outside of the upper range of the margin of error of the population. The mean for the Deaf and Paralyzed samples fall outside of the lower range of the margin of error of the population.

*Step 9: Plotting Subgroup Means with the Different Sampling Methods*

* Plot (on one graph) - 1) the computed population mean/standard deviation (Step 3), (2) the computed mean/standard deviation using the random sampling method for each subgroup when using 50% of the data (Step 6), and 3) the computed mean/standard deviation using the stratified random sampling method for each subgroup (Step 7). Which sampling method has the largest error when comparing to the population mean? Define your method for computing error.
  + Graph on next page.
  + Error:
    - Average error for Random Sampling method: 4.583389210256656
    - Average error for Stratified Random Sampling method: 5.208550500325312
    - The Stratified Random Sampling method had a larger error compared to the population mean. My method for computing this was to do for each subgroup within each sampling method and then average the three subgroups to find an average for each sampling method.
    - The higher error for Stratified Random Sampling method is easily seen in the graph below as those associated bars vary much more then the Random Sampling method bars.

