Who is your TA (check one)?

🞏 De Andre 🞏 Corey 🞏 Jingkai 🞏 Neeraj

**SOC 3811/5811 – STATA Assignment #1**

PART 1. FOR ALL STUDENTS IN SOC **5811**

Go the GSS Data Explorer website at: <https://gssdataexplorer.norc.org/>. You may need to register for a (free) account. Once you are logged in, go to “MyGSS” and create a new project. When prompted, specify that you want “Stata (Control + System)” files (and not files for other software packages). Your cart will automatically include three variables, including “year.” Add the variable “wordsum” to your cart. Wordsum is a 10-item vocabulary test. (Before going any further, please go read about the GSS’s WORDSUM test here: <https://en.wikipedia.org/wiki/Wordsum>.) Then, create a data extract that includes the variables “year” and “wordsum.”

Once the extract is ready, the GSS Data Explorer will provide Stata code for reading in the two variables. Be sure to declare missing values to be missing!Then do the following:

1. Create a table below with frequency and percentage distributions of wordsum (for all years, not for each separate year).

Value Freq Pct Cum Pct

 0 228 0.71 0.71

 1 587 1.84 2.55

 2 1,036 3.24 5.79

 3 1,914 5.99 11.78

 4 3,268 10.23 22.01

 5 5,285 16.54 38.55

 6 7,010 21.94 60.49

 7 5,091 15.93 76.43

 8 3,423 10.71 87.14

 9 2,440 7.64 94.78

 10 1,668 5.22 100.00

 Total 31,950 100.00

1. Create a table below that reports the mean, median, and standard deviation of wordsum scores for each GSS year.

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Mean** | **Median** | **SD** |
| 1974 | 6.0 | 6 | 2.2 |
| 1976 | 6.1 | 6 | 2.2 |
| 1978 | 6.0 | 6 | 2.2 |
| 1982 | 5.7 | 6 | 2.2 |
| 1984 | 6.0 | 6 | 2.2 |
| 1987 | 5.7 | 6 | 2.2 |
| 1988 | 5.8 | 6 | 2.1 |
| 1989 | 5.9 | 6 | 2.2 |
| 1990 | 6.1 | 6 | 2.2 |
| 1991 | 6.1 | 6 | 2.1 |
| 1993 | 6.0 | 6 | 2.2 |
| 1994 | 6.1 | 6 | 2.1 |
| 1996 | 6.0 | 6 | 2.1 |
| 1998 | 6.1 | 6 | 2.1 |
| 2000 | 6.0 | 6 | 2.1 |
| 2004 | 6.2 | 6 | 2.1 |
| 2006 | 6.2 | 6 | 2.0 |
| 2008 | 5.9 | 6 | 2.0 |
| 2010 | 6.0 | 6 | 2.1 |
| 2012 | 5.9 | 6 | 2.1 |
| 2014 | 6.0 | 6 | 2.0 |
| 2016 | 6.0 | 6 | 1.9 |
| 2018 | 5.9 | 6 | 2.0 |

1. Answer this question in words: How has the distribution of wordsum scores changes over the years?

It really hasn’t changed. The mean has been steady at about 6, the SD at about 2.

(This is surprising, by the way, since levels of education in American rose a lot over these years!)

1. Paste below the Stata syntax you used to read in the GSS extract and perform the analyses above.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* These commands tell STATA how to interpret things

\* 1. cd tells STATA what computer directory you are working in

\* 2. "set more off" tells STATA's results screen to scroll nicely

\* 3. the log commands tell the computer to save an output file with results

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

cd "C:\Users\warre046\Desktop"

set more off

log using "assignment1.log", replace

log on

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* These commands read the raw data into STATA

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

infix year 1-20 wordsum 21-40 using GSS.dat

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* These commands labels the variables

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

label variable year "Gss year for this respondent"

label variable wordsum "Number words correct in vocabulary test"

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* These commands label each variable's values

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

label define gsp001x 99 "Did not try" -1 "Not applicable"

label values wordsum gsp001x

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* These commands decalre some values to represent missing data

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

replace wordsum=. if wordsum==99

replace wordsum=. if wordsum==-1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* QUESTION #1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

tab wordsum

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* QUESTION #2

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

bysort year: tabstat wordsum, stat(mean median sd)

log off

log close

PART 2. FOR ALL STUDENTS IN SOC **3811**

I recently ran an internet survey of a random sample of 2,181 Americans. I collected information about each person’s age, education, race/ethnicity, and sex. I also asked them four questions about whether they had engaged in certain delinquent/deviant behaviors. The data that resulted from this survey are contained in the data file for this assignment. Below I include “meta-data” --- information about each variable, including (1) the name of the variable; (2) the variable’s column locations in the data file; (3) a brief description of the variable; and (4) the label associated with each numeric value. None of the variables have any missing data.

Use the example STATA syntax file (on the course web site) and modify it to accomplish the following goals. When you are done, type or paste your answers for questions #2 through #4 below into the Word file and turn it in via Canvas.

1. Read the data file into STATA
2. Produce frequency distributions for each of the variables for the entire sample.

|  |  |  |  |
| --- | --- | --- | --- |
| **Arrested** | Freq. | Percent | Cum. |
|  |  |  |  |
| Yes | 365 | 16.75 | 16.75 |
| No | 1,814 | 83.25 | 100 |
|  |  |  |  |
| **Pot** | Freq. | Percent | Cum. |
|  |  |  |  |
| Yes | 980 | 44.97 | 44.97 |
| No | 1,199 | 55.03 | 100 |
|  |  |  |  |
| **Theft** | Freq. | Percent | Cum. |
|  |  |  |  |
| Yes | 767 | 35.2 | 35.2 |
| No | 1,412 | 64.8 | 100 |
|  |  |  |  |
| **Drunk** | Freq. | Percent | Cum. |
|  |  |  |  |
| Yes | 1,146 | 52.59 | 52.59 |
| No | 1,033 | 47.41 | 100 |
|  |  |  |  |
| **Age** | Freq. | Percent | Cum. |
|  |  |  |  |
| 18 | 10 | 0.46 | 0.46 |
| 19 | 29 | 1.33 | 1.79 |
| 20 | 24 | 1.1 | 2.89 |
| 21 | 28 | 1.28 | 4.18 |
| 22 | 25 | 1.15 | 5.32 |
| 23 | 20 | 0.92 | 6.24 |
| 24 | 30 | 1.38 | 7.62 |
| 25 | 22 | 1.01 | 8.63 |
| 26 | 31 | 1.42 | 10.05 |
| 27 | 21 | 0.96 | 11.01 |
| 28 | 29 | 1.33 | 12.35 |
| 29 | 22 | 1.01 | 13.35 |
| 30 | 21 | 0.96 | 14.32 |
| 31 | 27 | 1.24 | 15.56 |
| 32 | 25 | 1.15 | 16.7 |
| 33 | 31 | 1.42 | 18.13 |
| 34 | 31 | 1.42 | 19.55 |
| 35 | 28 | 1.28 | 20.84 |
| 36 | 26 | 1.19 | 22.03 |
| 37 | 26 | 1.19 | 23.22 |
| 38 | 29 | 1.33 | 24.55 |
| 39 | 40 | 1.84 | 26.39 |
| 40 | 43 | 1.97 | 28.36 |
| 41 | 48 | 2.2 | 30.56 |
| 42 | 40 | 1.84 | 32.4 |
| 43 | 37 | 1.7 | 34.1 |
| 44 | 39 | 1.79 | 35.89 |
| 45 | 17 | 0.78 | 36.67 |
| 46 | 32 | 1.47 | 38.14 |
| 47 | 39 | 1.79 | 39.93 |
| 48 | 34 | 1.56 | 41.49 |
| 49 | 43 | 1.97 | 43.46 |
| 50 | 38 | 1.74 | 45.2 |
| 51 | 48 | 2.2 | 47.41 |
| 52 | 53 | 2.43 | 49.84 |
| 53 | 48 | 2.2 | 52.04 |
| 54 | 46 | 2.11 | 54.15 |
| 55 | 50 | 2.29 | 56.45 |
| 56 | 49 | 2.25 | 58.7 |
| 57 | 53 | 2.43 | 61.13 |
| 58 | 57 | 2.62 | 63.74 |
| 59 | 57 | 2.62 | 66.36 |
| 60 | 55 | 2.52 | 68.88 |
| 61 | 48 | 2.2 | 71.09 |
| 62 | 49 | 2.25 | 73.34 |
| 63 | 38 | 1.74 | 75.08 |
| 64 | 48 | 2.2 | 77.28 |
| 65 | 42 | 1.93 | 79.21 |
| 66 | 43 | 1.97 | 81.18 |
| 67 | 41 | 1.88 | 83.07 |
| 68 | 47 | 2.16 | 85.22 |
| 69 | 42 | 1.93 | 87.15 |
| 70 | 39 | 1.79 | 88.94 |
| 71 | 30 | 1.38 | 90.32 |
| 72 | 34 | 1.56 | 91.88 |
| 73 | 21 | 0.96 | 92.84 |
| 74 | 28 | 1.28 | 94.13 |
| 75 | 14 | 0.64 | 94.77 |
| 76 | 18 | 0.83 | 95.59 |
| 77 | 13 | 0.6 | 96.19 |
| 78 | 14 | 0.64 | 96.83 |
| 79 | 14 | 0.64 | 97.48 |
| 80 | 5 | 0.23 | 97.71 |
| 81 | 11 | 0.5 | 98.21 |
| 82 | 7 | 0.32 | 98.53 |
| 83 | 9 | 0.41 | 98.94 |
| 84 | 5 | 0.23 | 99.17 |
| 85 | 4 | 0.18 | 99.36 |
| 86 | 7 | 0.32 | 99.68 |
| 87 | 3 | 0.14 | 99.82 |
| 88 | 1 | 0.05 | 99.86 |
| 90 | 1 | 0.05 | 99.91 |
| 92 | 1 | 0.05 | 99.95 |
| 93 | 1 | 0.05 | 100 |
|  |  |  |  |
| **Education** | Freq. | Percent | Cum. |
|  |  |  |  |
| No formal education | 4 | 0.18 | 0.18 |
| 1st, 2nd, 3rd, or 4th grade | 1 | 0.05 | 0.23 |
| 5th or 6th grade | 2 | 0.09 | 0.32 |
| 7th or 8th grade | 27 | 1.24 | 1.56 |
| 9th grade | 17 | 0.78 | 2.34 |
| 10th grade | 31 | 1.42 | 3.76 |
| 11th grade | 31 | 1.42 | 5.19 |
| 12th grade NO DIPLOMA | 53 | 2.43 | 7.62 |
| HIGH SCHOOL GRADUATE - high school DIPL | 657 | 30.15 | 37.77 |
| Some college, no degree | 456 | 20.93 | 58.7 |
| Associate degree | 177 | 8.12 | 66.82 |
| Bachelors degree | 425 | 19.5 | 86.32 |
| Masters degree | 219 | 10.05 | 96.37 |
| Professional or Doctorate degree  | 79 | 3.63 | 100 |
|  |  |  |  |
| **Race** | Freq. | Percent | Cum. |
|  |  |  |  |
| White, NH | 1,690 | 77.56 | 77.56 |
| Black, NH | 185 | 8.49 | 86.05 |
| Other, NH | 81 | 3.72 | 89.77 |
| Hispanic | 160 | 7.34 | 97.11 |
| 2+ Races | 63 | 2.89 | 100 |
|  |  |  |  |
| **Sex** | Freq. | Percent | Cum. |
|  |  |  |  |
| Male | 1,118 | 51.31 | 51.31 |
| Female | 1,061 | 48.69 | 100 |

1. Produce frequency distributions of the four deviance/delinquency variables separately for men and women

|  |  |  |
| --- | --- | --- |
|  | **Men** | **Women** |
| **Arrested** |  |  |
|  |  |  |
| Yes | 282 | 83 |
| No | 836 | 978 |
|  |  |  |
| Total | 1,118 | 1,061 |
|  |  |  |
|  |  |  |
| **Pot** |  |  |
|  |  |  |
| Yes | 563 | 417 |
| No | 555 | 644 |
|  |  |  |
| Total | 1,118 | 1,061 |
|  |  |  |
|  |  |  |
| **Theft** |  |  |
|  |  |  |
| Yes | 477 | 290 |
| No | 641 | 771 |
|  |  |  |
| Total | 1,118 | 1,061 |
|  |  |  |
|  |  |  |
| **Drunk** |  |  |
|  |  |  |
| Yes | 715 | 431 |
| No | 403 | 630 |
|  |  |  |
| Total | 1,118 | 1,061 |

1. Answer these questions based on the results above:
	1. What percentage of people did not complete high school? **7.62%**
	2. What percentage has ever smoked pot? **45%**
	3. On which of the four deviance/delinquency variable are men’s and women’s answers most different? That is, for which of the four do men and women give the most divergent answers? **Drunk**