Sociology 2200: Statistics
Instructor: Natasha Sarkisian

Assignment 1 Answers

This assignment is focused on using some basic statistical terminology and on calculating and interpreting descriptive statistics, by hand and using Stata. It relies on the course material addressed in modules 1 and 2.

Part I. An article in the June 10, 2002, issue of the Archives of Internal Medicine reported on a study of the effectiveness of a nicotine lozenge for helping smokers to quit smoking. Smokers who participated in the study were randomly assigned to receive either the nicotine lozenge being tested or a placebo (no active ingredient) lozenge. Several characteristics were recorded on each subject at the beginning of the study, such as weight (in lbs), gender, number of cigarettes smoked per day, and whether or not the person had made any previous attempts to quit smoking. At the end of the study, the study assessed whether or not the subject had successfully refrained from smoking.

1. List all the variables mentioned in this study description and specify their level of measurement.

Variable	**Level of measurement**
Nicotine vs placebo group	Nominal
Weight	Ratio
Gender	Nominal
Number of cigarettes smoked per day	Ratio
Previous attempts to quit smoking	Nominal
Successfully refrained from smoking	Nominal

2. What are the units of observation in this study?
   Individual smokers

3. Most likely, what is the <u>main</u> research question in this study?

Does using nicotine lozenges increase the likelihood that a person will be able to refrain from smoking?

Part II. Assume quiz scores for 10 people (variable X) are: 5, 5, 9, 6, 7, 7, 8, 6, 8, 6.

1. Organize the data – build a table of values, then calculate the frequencies. Express the frequencies in percentages. Calculate cumulative percentages.

Value of X	Frequency	Percent of cases	Cumulative percent
5	2	20%	20%
6	3	30%	50%
7	2	20%	70%
8	2	20%	90%
9	1	10%	100%

2. Find the mean using the frequency distribution to facilitate addition.

Mean=(5*2 + 6*3 + 7*2 + 8*2 + 9*1)/10=67/10=6.7

3. Find the median using the sorted list; then do it using the cumulative distribution.

Sorted list: 5, 5, 6, 6, 6,| 7, 7, 8, 8, 9
Median=(6+7)/2=6.5
Cumulative distribution: 6.

4. Find the mode.

Mode=6 (30%)

5. Find the variation ratio.

Subtracting the percentage in modal category from 100%: 100%-30%=70% is variation ratio.

6. Find the range.

Range=max-min=9-5=4

7. Find the interquartile range using both the sorted list approach and the cumulative distribution approach.

Sorted list: 5, 5, 6, 6, 6,| 7, 7, 8, 8, 9
Q3=8
Q1=6
ICQ=Q3-Q1=8-6=2

Cumulative distribution: Q1=6, Q3=8, IQR=8-6=2

8. Find the standard deviation and the variance.

X	$X-\bar{X}$	$(X-\bar{X})^2$
5	-1.7	2.89

5	-1.7	2.89
9	2.3	5.29
6	-0.7	0.49
7	0.3	0.09
7	0.3	0.09
8	1.3	1.69
6	-0.7	0.49
8	1.3	1.69
6	-0.7	0.49
sum	0	16.1

or

Value of X	Frequency	$X - \bar{X}$	$(X - \bar{X})^2$	$(X - \bar{X})^2 *$ Frequency
5	2	-1.7	2.89	2.89*2=5.78
6	3	-0.7	0.49	0.49*3=1.47
7	2	0.3	0.09	0.09*2=0.18
8	2	1.3	1.69	1.69*2=3.38
9	1	2.3	5.29	5.29*1=5.29
sum				16.1

Variance = 16.1/9=1.79
Standard deviation = sqrt(1.79)=1.34

9. In words, what does the number that you calculated for standard deviation tell us.
It tells us that on average, the values in this distribution are 1.34 points away from the mean. In other words, a typical distance between an observation and the mean is 1.34 in this set of data.

10. Calculate $(X - \bar{X})$ for the third score, X=9. In words, what does this number mean?

9-6.7=2.3
That number shows the distance between the given value and the mean -- that is, the third score is 2.3 units above the mean.

Part III: Stata

Using the General Social Survey 2012 data (gss2012.dta, available from Canvas), conduct the following analyses and answer the questions below. Make sure to open a log file before starting (use .log extension). Please do not submit screenshots or copy/paste from the screen – I need your actual log file.

Construct a frequency distribution and obtain measures of central tendency (mean, median, mode) and variability (range, interquartile range, standard deviation, variance) to describe the variable *educ*. Fill the blanks and answer the questions below based on these data.

1. On average, the mean number of years of education people have is **_13.53____**.

2. The median number of years is **__13___**.

3. The mode for the number of years of education is **___12___**.

4. Which measure of central tendency do you think best describes this distribution, and why?

There can be different opinions here—all three could be used in principle (as long as the reasoning is right), and they are not very far from each other. Median and mean are close to each other and both are good measures here because it is a ratio-level variable. The distribution is not highly skewed, and outliers are not particularly influential here, so mean might be the best choice. Mode is usually not selected for ratio-level variables, but here it could still be a good choice because education can also be considered ordinal (because 1 year of education has different meaning since some levels, like 12, correspond to completed degrees). Mode in this case highlights the most frequent category in the sample that is very common (that category is those who completed high school but didn't attend any college).

5. The standard deviation for the number of years of education is **__3.1_____**.
What does this number tell us (in words)?
On average, individuals' education diverges from the mean by approximately 3.1 years.

6. The <u>least</u> common number of years of education in this sample is **_1_____**.

```
------------------------------------------------------------------------------
      log:  A1.log
  log type:  text
 opened on:  20 Sep 2025, 20:58:59

. tab educ

    HIGHEST |
    YEAR OF |
     SCHOOL |
  COMPLETED |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          3        0.15        0.15
          1 |          2        0.10        0.25
          2 |          3        0.15        0.41
          3 |          6        0.30        0.71
          4 |         10        0.51        1.22
          5 |          4        0.20        1.42
          6 |         27        1.37        2.79
          7 |          8        0.41        3.19
          8 |         48        2.43        5.63
          9 |         47        2.38        8.01
```

```
         10 |         59        2.99       11.00
         11 |        101        5.12       16.13
         12 |        540       27.38       43.51
         13 |        163        8.27       51.77
         14 |        261       13.24       65.01
         15 |         99        5.02       70.03
         16 |        307       15.57       85.60
         17 |         80        4.06       89.66
         18 |         92        4.67       94.32
         19 |         41        2.08       96.40
         20 |         71        3.60      100.00
------------+-----------------------------------
      Total |      1,972      100.00

. tabstat educ, stats(mean median range iqr sd variance)

   variable |      mean       p50     range       iqr        sd  variance
------------+--------------------------------------------------------------
       educ |  13.52789        13        20         4  3.126576  9.775477
------------------------------------------------------------------------------

. log close
      log:  A1.log
  log type:  text
 closed on:  10 Sep 2024, 21:03:32
------------------------------------------------------------------------------
```