K300 (4392) Statistical Techniques (Fall 2007)

Assignment 2: Descriptive Statistics (Due September 17)

Instructor: Hun Myoung Park
kucc625@indiana.edu, (317) 274-0573

You should always read instructions carefully before solving questions.

This assignment contains exercises for summarizing univariate variables using graphical and numerical methods (total 140 points). Please read the instructions carefully to do the homework successfully.

- You MAY NOT use a wordprocessor (e.g., Microsoft Word and WordPerfect); **Use separate sheets and write down answers by hand.** Exception is SPSS outputs for questions 12-14.
- Follow all instructions of a question. Do not skip any one.
- Try to show how you obtain the answers (statistics). In some questions, a single number may not be accepted as the answer.
- Useful examples and sections of the textbook are italicized for you.
- Put your answer sheets into an envelope and hand in on Monday, September 17.
- You MAY NOT discuss with other classmates when answering the questions.

If you have any problem with any of the questions, please post messages on Oncourse CL or come and see the instructor during office hour MW 2:00-3:00 P.M. Or you may make an appointment with the instructor.

1. (10 points) Classify the level of measurement (nominal, ordinal, interval, and ratio) of the following variables. Note that some variables may be used as more than one level of measurement depending on research questions and data at hand. *See page 6-9 and question 7 on page 26*.
You need to ask yourself such questions as “Is there any meaningful rank order between two values?” “Is there any meaningful difference between two variables? “Is there infinite numbers between two particular values? (continuous or discrete)” and “Can I apply arithmetic operators (addition, subtraction, multiplication, and division) to values of a variable?” And then carefully evaluate the extent that the variable conveys information.

1) What was your GPA last semester? (interval)
2) What kind of beer do you prefer? (nominal) Is Miller Light larger than Coors Light? Can you subtract Miller Light from Coors Light? How about Miller light divided by Coors Light? If I ask “To what extent do you like Miller Light?”, it may be ordinal or interval, depending on the specific format of answer. Do not be confused by the word “prefer.”
3) Where are you from? (nominal) Can you imagine Indiana times Ohio? What do you think about China = Korea + Japan × U.S.? Does that make sense? Is there any meaningful rank order between the two states? If you argue Indiana is smaller than Ohio,

you may not be saying names of the states but comparing areas of two states. The name and area are aspects (attributes or characteristics) of a state. But they are different variables. Name is nominal, where area is interval/ratio. Similarly, Indianapolis is larger than Bloomington in terms of several aspects (e.g., area and population), but the names themselves are not comparable. Do not argue like "Indianapolis has 12 characters, 1 more than Bloomington. So Indianapolis > Bloomington." You are not talking about the name of a city but the length of a city name; they are totally different variables. "Where are you from?" does not ask the population and size of your home state. Therefore, you may not answer as "5 millions" or "smaller than California." The question simply asks the name of your home state (or country).

4) What do you think about Indiana property tax increase? (Strongly agree—Agree—Indifference—Disagree--Strongly disagree) (ordinal in general) Probably interval if assuming the equal interval between two categories. For example, difference between "Strongly agree" and "agree" is assumed to be equal to that of "Indifference" and "Disagree." In general, the former is considered larger than the latter.

5) Age (<=20; <=25; <=30; <=40; <=50; <=60) (ordinal but interval as well)
6) How many credit hours did you take last semester? (interval)
7) Grade (from 1st through 12th) (ordinal, interval if you consider grade as the number of years that a person studied in schools)
8) Fahrenheit temperature at 9:00 A.M. in IUPUI (interval) Never ratio. 0 degrees Fehrenheit does not mean an object does not have heat. Note that even ice has some heat. Kelvin's 0K (or -459.67F) means no heat in the object.
9) Your monthly gas consumption (gallons) (interval/ratio)
10) which one (1 through 9), if any, can be treated as more than one level of measurement? And why? You may make your own assumptions to support your arguments, if needed. (4, 5, 7)

2. (10 points) Explain what the stratified sampling is and then apply it to a study on student satisfaction at IUPUI/SPEA. Given different numbers of students enrolled in each major/department this semester, present your strategy to obtain a representative sample. Ignore other variables such as gender and race. Sample size is limited to 500. *See page 12 and question 12 on page 26.*

The stratified sampling divides the population into strata according to some characteristics in order to make the sample more representative, and then randomly draws sample from each stratum. In this study, we want to take 500 samples from majors/departments in proportion to their number of enrolled students. Thus, I would divide IUPUI/SPEA students into individual majors/departments and check the proportion of a major of the total enrolled students. And then draw random samples from each stratum. If a policy major accounts for 20percent of total enrolled students, I would take 100 random samples (=500*.20) from the major. I don't mind how many female or Hispanic students are selected from each major because this is not my concerning in this case.

3. (10 points) Solve question 9 on page 44. You need to construct a frequency table with five classes and compute the mean on the basis of the frequency table you construct. And then compute the mean of raw data. Which one is larger? *See pages 37-41 for examples*.
First, I need to decide the class width for five classes. The range is 13 = 32-19. The class width is 2.6 = 13/5. So the first class is from 19.0 through 21.6 (=19.0+2.6) exclusive. The second class includes values greater than or equal to 21.6 and smaller than 24.2 (=21.6+2.6). And so on. I would count frequencies of data points falling into corresponding classes. For example, I found 2 data points that greater than or equal to 19.0 and smaller than 21.6.

In order to compute mean from the frequency table, I need to know the midpoint of each class. Midpoint, by definition, is the mean of two cut points. For example, the midpoint of the first class is 20.3=(19.0+21.6)/2. And then I multiply the frequency and midpoint of each class. For instance, I get 298 (=13 X 22.9) for the second class. When summing all multiplications up, I get 736 (=41+298+…31). Now, I divide 736 by the number of observations 30 to get the mean of 24.5467.

Class Limits	Frequency	Midpoint	Frequency * Midpoint
19.0-21.6	2	20.3	41
21.6-24.2	13	22.9	298
24.2-26.8	10	25.5	255
26.8-29.4	4	28.1	112
29.4-32.0	1	30.7	31
Sum	30		736

Mean 24.54666667 =736/30

You may feel difficulty figuring out this computation. Why do we need to compute midpoint? This is because we lose some information when classifying raw data into this frequency table. (Summarization inevitably involves information loss to some degree.) Therefore, we need to get a value that **represents** the class. You may want to use the low bound of each class like 19.0 and 21.6, but midpoints appear more reasonable. Don't you think so? Once you get the midpoint, the computation will be easy if you think a bit differently. How? (20.3+20.3) + (22.9+22.9+22.9…) + (25.5+25.5+…) + (28.1+28.1+28.1+28.1) + (30.7) = (20.3 X 2) + (22.9 X13) + (25.5 X 10) + (28.1 X 4) + (30.7 X 1) = 736. What do you think about it? This is simple summation. No puzzle, no mystery.

Now, let us sum all raw data up. I get the sum of 737, 1 larger than the sum from the frequency table. If you use upper bound of each class, for example, the sum from the frequency table must be much larger than 737. Thus, midpoint must be a best choice.
I get the mean 24.5667 (= 737/30), which is slightly large than the mean based on the frequency table.

Keep in mind you may lose some information when constructing a frequency table. On the other hand, a frequency table provides a big picture of data at hand that raw data themselves cannot.

4. (10 points) Solve question 14 on page 44. And draw a stem-and-leaf plot with 8 stems (leading digits) by hand. *See pages 73-76.*

The data have a wide range: 3970=4040-70. In order accommodate all data points in 8 classes, I would begin with 70 with a class width of 500. Why 500? 3970/8 is 496.25 but 500 is much easier to use for computation. You may set the beginning point other than minimum (70) or class width as long as classes can accommodate all data points. (For example, 50-546.25, 546.25-1042.6… Do you like this scheme?) In other ward, if your beginning point and class width, whatever you set, fail to cover all data points, you are off the track. Keep in mind, class width should be equal across classes. Why? Otherwise, a frequency table may exaggerate the "big picture" because class width influences frequencies in individual classes. To sum, my principle for constructing a frequency table is 1) all data points are included in any one class (you may adjust beginning point and class width), 2) each class has the same class width, 3) class limits (or boundaries) should be easy to handle or read (Do not use "75.25424-575.2456" and the like), 4) any class may not skipped regardless of the number of frequency of the class.

So my first class begins with 70 and ends with 570. I found 14 data points falling into the range. This 14 accounts for 52 percent (14/27). I found 5 data points greater than or equal to 570 and smaller than 1070, and its relative frequency is 19 percent (5/27). The cumulative relative frequency in the second class is .70 (=.52+.19).

Class Limits	Frequency	Cumulative frequency	Relative frequency	Cumulative relative frequency
70-570	14	14	0.52	0.52
570-1070	5	19	0.19	0.70
1070-1570	5	24	0.19	0.89
1570-2070	0	24	0.00	0.89
2070-2570	0	24	0.00	0.89
2570-3070	1	25	0.04	0.93
3070-3570			0.00	0.93
3570-4070	2	27	0.07	1.00
Sum	27		1	

In order to draw a stem-and-leaf plot, I sort the data in the ascending order. Given a wide range, the leading digit appears messy. My principles are 1) each stem has the same length, 500 in this case, 2) leading digits should be meaningful; you should be able to replicate original data point using the stem-and-leaf plot, 3) any stem should not be

skipped even when the stem does not have any data point. You may not use serial numbers like 1, 2, 3, 4, 5…, 3) You may not omit any stem in the middle regardless of the number of data points falling into the stem. Why? Omitting exaggerates the distribution of data you have. Think about what the following plot will look like if you omit the two stems that do not have and data points. The skewness may look less sever that should be.

0*** | 70, 85, 125, 145, 180, 205, 260, 300, 325, 350, 405, 460, 480, 485
0*** | 620, 690, 705, 875, 970
1*** | 100, 160, 430, 430
1*** | 555
2*** |
2*** | 805
3*** |
3*** | 630
4*** | 040

Let us take a look at the second stem of the plot as an example. The "***" of the leading digit "0***" should be replaced by individual tailing digits on the right. So 620 is 0620, 690 is 0690, and so on. Similarly, the four stem has a data point of 1555. That is the reason why I say the leading digit should be meaningful. In this example, there are two or three digits in the right-hand side, which need a comma to separate data points. If you have only one digit, you do not need to use a comma.

You have to recognize that the stem-and-leaf plot shows the (probability) distribution of data and contains raw data without any information loss (you can replicate raw data from the stem-and-leaf plot). Of course, this plot has shortcomings as well. It is not easy to draw especially when you have many data points. How do you manually draw the stem-and-leaf plot for 1 million data points? How do you sort them in the ascending or descending order? You may need many papers or huge blackboard in addition to pencils and erasers. I would give up and go home!

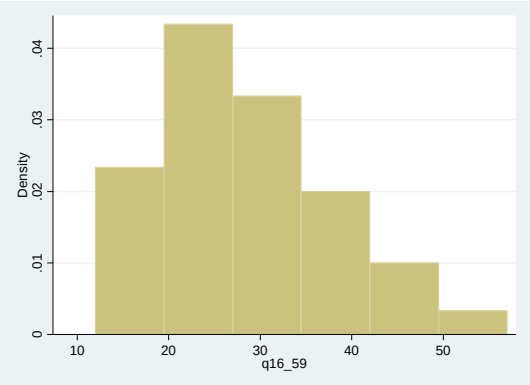
5. (10 points) Solve question 16 on page 59. You may skip the frequency polygon and ogive for the data. *See Figure 2-8 on page 56 for description of the histogram.*

The range is 45 = 57-12. Class width is about 7.5 = 45/6. I would use 8 instead for convenience. Depending on class width and class limits, frequencies may vary. See question 3 for constructing a frequency table.

Class Limits	Frequency	Relative frequency	Cumulative relative frequency
10.0-18.0	5	0.13	0.13
18.0-26.0	12	0.30	0.43
26.0-34.0	12	0.30	0.73
34.0-42.0	7	0.18	0.90
42.0-50.0	3	0.08	0.98

50.0-58.0	1	0.03	1.00
Sum	40	1.00	

The histogram appears right skewed.



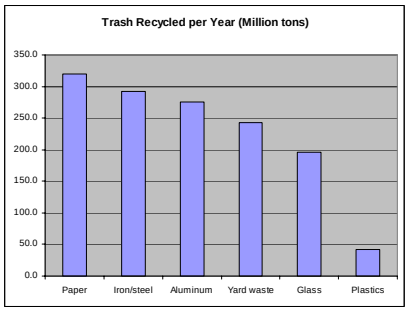
6. (10 points) Solve question 18 on page 79. Keep in mind that whole numbers are nonnegative integers (i.e., 0 and positive integers). You may have 10 stems (leading digits). *See example 2.14 on page 76*.

This stem-and-leaf plot makes it easy to compare two distributions. Leading digits begins from 0. and ends with 9. The tailing digit has only one number and thus a comma is not necessary. For example, the second stem reads 1.5 and 1.9. You should not omit any stem. Do not forget to sort data in ascending order.

Females		Males
5	0.	3
	1.	59
	2.	2
74320	3.	11
6	4.	1466
9630	5.	26669
85	6.	0066
720	7.	7
876600	8.	78
42	9.	68

The distribution of male appears symmetric, while female data tend to be distributed randomly without any pattern.

7. (5 points) Solve question 24 on pages 97-98. Draw a Pareto chart by hand. *See pages 64-65*.



This is a typical bar chart. You may have a similar Pareto chart for the data.

8. (5 points) Solve question 12 on page 117. Simply compute the mean only. You SHOULD show each step you follow as showing in *the example 3-3 on pages 105-106*.

N=108, mean is 113=12204/108. See question 3 as well.

Class Limits	Frequency	Midpoint	Frequency * Midpoint
90-98	6	94	564
99-107	22	103	2,266
108-116	43	112	4,816
117-125	28	121	3,388
126-134	9	130	1,170
Sum	108		12,204
Mean	113	(=12,204/108)	

9. (5 points) Solve question 15 on page 117. You do not need to compute the mean in order to answer. *See "Properties and Uses of Central Tendency" on page 114.*
Probably not. Data points are highly skewed to the right or concentrated on the left. As a result, the mean may not be a good measure that summarizes the data. Median (or mode in this case) will be much better.

10. (10 points) Solve question 1 on page 163. Add "25" to the series of data so that you have 8 data points. Obtains key descriptive statistics (minimum, 1Q, 2Q, 3Q, maximum, and mean) and draw the box plot based on the statistics. A box plot may be arranged either horizontally or vertically (either one will do). And then mark all data points using "O" on the box plot. Put "X" on the mean position of the box plot. Is the mean larger than the median? *See pages 147-149 and 159-161 for obtaining the key statistics*.

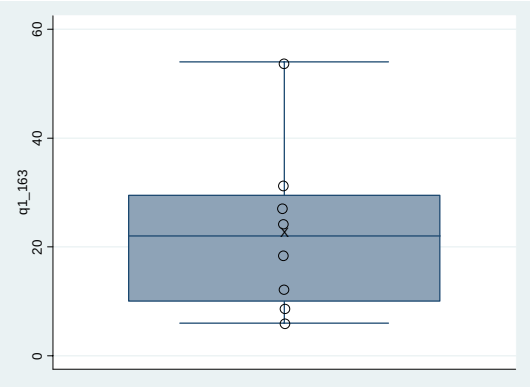
Ordered data: 6, 8, 12, 19, 25, 27, 32, 54

```
. tabstat q1_163, stat(min, p25, p50, p75, max, mean)

    variable |       min       p25       p50       p75       max      mean
-------------+--------------------------------------------------------------
      q1_163 |         6        10        22      29.5        54    22.875
--------------------------------------------------------------------------
```

Minimum: 6
1Q: 10 = (8+12)/2
2Q (median) : 22 = (19+25)/2
3Q: 29.5 =(27+32)/2
Maximum: 54

My box plot is vertically arranged but your may be horizontally arranged.



11. (15 points) Solve question 7 on page 135 but compute the mean, variance, and standard deviation ONLY. You SHOULD show steps for computation as we did on an Excel sheet during the lab. DO NOT use Excel though; DO it by hand. *See examples 3-21 (step 6) and 3-22 (step 3) on pages 123-125.* Keep in mind the data points were drawn from the population; you need to compute sample variance and standard deviation (as opposed to the population counterparts).

See the excel sheet that we discussed in the first lab.
N: 10
Sum: 133 = 7+37+…+3
Mean: 13.3=133/10
Sum of (y-y bar)^2=2292.1
Variance: 254.6778=2292.1/(10-1) You should divide by n-1 because this is a sample.
Alternatively, 254.6778=[4061-(133^2)/10]/(10-1)
Standard deviation: 15.9586 =sqrt(254.6778)

Series	Data	Y-mu	(y-mu)^2	y^2
1	7	-6.3	39.69	49
2	37	23.7	561.69	1369
3	3	-10.3	106.09	9
4	8	-5.3	28.09	64
5	48	34.7	1204.09	2304
6	11	-2.3	5.29	121
7	6	-7.3	53.29	36

	8	0	-13.3	176.89	0
	9	10	-3.3	10.89	100
	10	3	-10.3	106.09	9
Sum		133	0	2292.1	4061

Mean	13.3	= (133/10)
Variance	254.6777778	= 2292/(10-1)
Std Deviation	15.95862706	=sqrt(254.678)

The data show a couple of odd observations that increases the variance. The standard deviation is larger than the mean.

* Now let us turn to our data set. Download the SPSS data set of class survey from OnCourse or course webpage at http://www.masil.org/method/statistics.html. If you need questionnaire and data dictionary, download http://www.masil.org/teach/k300/K300_Survey.pdf.
Launch SPSS 15 for Windows and read the data set stored in your computer.

12 (15 points) Draw a frequency table of `major`. Click Analyze→Descriptive Statistics→Frequencies…. Choose `major` and click arrow to move the variable into the right. Click "Statitics" button on the bottom and check mean, median, variance, and "std deviation." Click Continue and then OK to get a frequency table. 1) Which major has the smallest and largest frequency? Management/Policy (smallest) and Criminal Justice (largest) 2) What is the proportion (relative frequency) of Health Administration of the total? How can you calculate by hand? 27.3=6/22*100 or 28.6=6/21*100 (excluding missing) 3) What are the mean and variance of `major`, 2.95 and 1.05 4) How do you interpret the mean and variance substantively? This question may remind you of the level of measurement. They are not interpretable. Why? This variable contains names of majors. There is no meaningful rank order and difference between two names. We may not divide "Criminal Justice" by "Health Administration." Obviously, this variable is nominal. Any arithmetic operation is not appropriate for categorical variables. Although we assign some numbers to major names for convenience, we cannot interpret substantively mean and variance, which are outcomes of arithmetic operation. 5) Report the median. Can you interpret the median substantively? 3. This median may contain certain information than mean, but not interpretable. This number should not be interpreted like "there must be many criminal justice students in data." 6) print out the SPSS output and append it to your assignment 2.

Statistics

major

N	Valid	21
	Missing	1
Mean		2.9524
Median		3.0000
Std. Deviation		1.02353
Variance		1.048

major

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Management/Policy	1	4.5	4.8	4.8
	Health Administration	6	27.3	28.6	33.3
	Criminal Justice	9	40.9	42.9	76.2
	Leadership	3	13.6	14.3	90.5
	Environment	2	9.1	9.5	100.0
	Total	21	95.5	100.0	
Missing	System	1	4.5		
Total		22	100.0		

13 (10 points) Produce descriptive statistics of `credits`. Click Analyze→Descriptive Statistic→Descriptives… and then choose `credits`. Click Options… on the right bottom, check sum, variance, and range, and then click Continue. 1) Print out the SPSS output and attach it to your assignment. 2) Report N, sum, and mean. 21, 285, 13.57 3) Show how the mean was calculated using these statistics. 13.57=285/21 4) Show how the standard deviation was calculated from the variance. 2.34=sqrt(5.45)

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
credits	21	10.00	6.00	16.00	285.00	13.5714	2.33605	5.45
Valid N (listwise)	21							

14 (15 points) Draw a histogram of `distance`. Click Analyze→Descriptive Statistic→Explore… Choose `distance` and click arrow so that the variable appears on a box under "Dependent List:" Click the "Statistics…" option at the bottom, check Percentiles, and click Continue. Now click "Plots" next to "Statistics…", check Histogram, and then click Continue. Finally click OK to get descriptive statistics and the histogram. 1) Print out SPSS output and attach to your assignment. 2) Report the mean and variance. 19.35 and 419.72 3) Report the key statistics used in a box plot. 1, 9 (1Q), 15 (2Q), 23.5 (3Q), 100 4) Look at the box plot. Write down five key statistics and the mean at their right positions on the box plot. Do you have any outliers (extremely large or small)? Yes. Probably 100 can be considered an outlier. You should write down these statistics on the plot. 5) Look at the histogram and stem-and-leaf plot. Is the histogram similar to the stem-and-leaf plot? If not, what do you think makes a difference? Note that graphs and charts oftentimes are sensitive to scale, number of classes, and outliers. Tell me your idea about difference and similarity of the histogram and stem-and-leaf plot of `distance`. The histogram and stem-and-leaf plot of this variable look different mainly due to use of different classes. Histogram may be misleading. 6) Finally, which statistics (mean or median) are you going to report as a representative value (central tendency) of `distance`? Which plot (histogram, box plot, stem-and-leaf plot) do you think will be best for summarizing this variable? And why? The presence of outliers implies that the

distribution is not symmetric and mean is not a good measure of central tendency. As a result, the median appears better than the mean in terms of a representative value. I would report the box-plot, which reports key statistics including outliers in an efficient way. The histogram may be misleading. The stem-and-leaf contains all information that the raw data have but does not provide numerical information such as mean, median, percentiles. This plot becomes problematic when N is extremely large; constructing and interpreting the stem-and-leaf plot must be a nightmare. The box plot is very informative in that it illustrate quartiles and outliers in an efficient way. I would choose the box plot or stem-and-leaf plot rather than the histogram for this varialbe.

Case Processing Summary

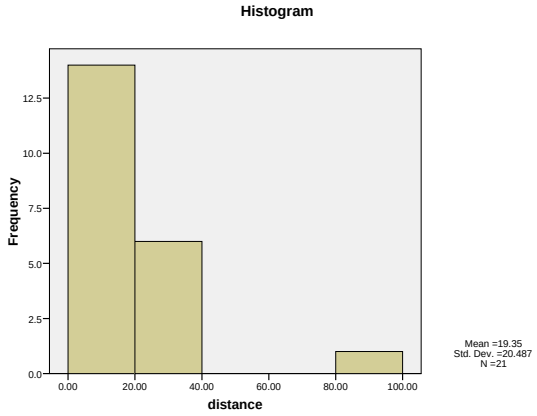
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
distance	21	95.5%	1	4.5%	22	100.0%

Descriptives

			Statistic	Std. Error
distance	Mean		19.3476	4.47063
	95% Confidence Interval for Mean	Lower Bound	10.0220	
		Upper Bound	28.6732	
	5% Trimmed Mean		16.0688	
	Median		15.0000	
	Variance		419.718	
	Std. Deviation		20.48701	
	Minimum		1.00	
	Maximum		100.00	
	Range		99.00	
	Interquartile Range		14.50	
	Skewness		3.254	.501
	Kurtosis		12.865	.972

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	distance	1.1000	2.0600	9.0000	15.0000	23.5000	30.0000	93.0000
Tukey's Hinges	distance			9.0000	15.0000	22.0000		



distance Stem-and-Leaf Plot

Frequency	Stem &	Leaf
3.00	0 .	122
3.00	0 .	699
1.00	1 .	0
7.00	1 .	5555578
2.00	2 .	02
1.00	2 .	5
3.00	3 .	000
1.00	Extremes	(>=100)

Stem width: 10.00
Each leaf: 1 case(s)