Exercise 1: Statistics 213 (L05) - Fall 2007

1. The universe or totality of items under consideration in a statistical study is a (sample, population, parameter, statistic)

2. the process of using information from a sample to draw conclusions about the entire population is called (sampling, the scientific method, statistical inference, descriptive statistics)

3. For each of the following situations describe the population of interest and the sample

   a. a politician who is running for the office of mayor of a city with 25000 registered voters commissions a survey which interviews 200 registered voters

   b. a manufacturer of computer chips claims that less than 10% of his products are defective. In order to check this claim 1000 chips were drawn from the population process and checked for defects.

4. For each question, identify each of the types of data requested as either quantitative or qualitative. When quantitative data is required, identify the variable as discrete or continuous.

   a. what is the weight of you car?

   b. in what city was your car made?

   c. how many cars are in your household?

   d. what is the distance traveled from your home to your school?

   e. what is the color of your car?

   f. what is the length of your car?

   g. what fuel economy (litres per 100km) do you get in city driving?

5. For a sample data: 87, 109, 79, 80, 96, 95, 90, 92, 96, 98, 101, 91, 78, 112, 94, 98, 94, 107, 81, 96, where $\sum x_i = 1874$, $\sum x_i^2 = 177328$, and $n = 20$

   a. Find mean, median, standard deviation, inter quartile range (IQR).

   b. Draw a box-plot.

   c. Determine the largest value is an outlier or not.

6. In 1789 the English scientist Henry Cavendish measured the density of the earth with great care. He repeated his measurement 28 times to obtain the following measurements for the density of the earth as multiple of the density of water: 5.50 5.57 5.42 5.61 5.53 5.47 5.62 5.63 5.07 5.29 5.34 5.26 5.44 5.46 5.55 5.34 5.30 5.36 5.79 5.75 5.29 5.10 5.68 5.58 5.27 5.85 5.65 5.39

   a. Draw a stem-and-leaf plot for this data

   b. Find mean, variance, standard deviation and median ($\sum x_i = 153.11$ and $\sum x_i^2 = 838.27$).

   c. Set up a frequency distribution table. Use a class width of 0.15 and take 5.0 for the lower boundary of the first class.

   d. Describe the shape of the distribution (is it symmetric, skewed to the left or right?) Are there obvious outliers ?

e. Would the mean and standard deviation be suitable measures to use to represent the center and the spread respectively of this distribution ? Would the median and interquartile be more appropriate ? Give reason for your answers.

f. Would you use the empirical rule ? Why or why not?

7. The 1993 New York Mets had the worst won-lost record in major league baseball. Here are the Mets' salaries, in thousands of dollars. 6200 5917 4000 3375 3000 2312 2300 2150 2100 1500 1012 850 650 635 500 475 220 205 195 195 158 145 109 109 109 109 109 ($\sum x_i = 38639, \sum x_i^2 - (\sum x_i)^2/n = 79783949.85$

a. If you were interested only in the New York Mets salaries in 1993, does this data represent a sample or a population ? Justify your answer.

b. Find the mean, median, variance and standard deviation.

c. Why might a stem-and-leaf plot not be as useful for this data as for the data in (1) ?

d. Using the following frequency distribution table, draw the relative frequency histogram.

salary ($000)	freqency	cumulative frequency	relative frequency
100 → 1000			
1000 → 1900			
1900 → 2800			
2800 → 3700			
3700 → 4600			
4600 → 5500			
5500 → 6400			
total			

e. Describe the shape of the distribution.

f. Would the mean and standard deviation be suitable measures to use to represent the center and the spread respectively of this distribution ? Would the median and interquartile be more appropriate ?

g. Would you use the empirical rule ? Why or why not ?

Partial solutions:

(3b). **population**: all computer chips produced and that will be produced in future by the manufacturer; **sample**: the 1000 chips checked

(5a). mean= 93.7, median= 94.5, sd= 9.55, position of $Q_1 = 5$, $Q_1 = 87$, position of $Q_3 = 16$, $Q_3 = 98$, $IQR = 11$

(5c.) $Z = \frac{112-93.7}{9.55} = 1.916 < 3$: it is not an outlier.

(6a). leaf unit (LU) =0.01; stem unit (SU)=0.1

stem	leaf
5.0	7
5.1	0
5.2	6 7 9 9
5.3	0 4 4 6 9
5.4	2 4 6 7
5.5	0 3 5 7 8
5.6	1 2 3 5 8
5.7	5 9
5.8	8

(6b). mean= 5.468, variance= 0.038, standard deviation= 0.196, median= 5.465

(6c). 5.0 → 5.15 includes 5.0, but not 5.15 and 5.15 → 5.30 includes 5.15, but not 5.30 ...

salary ($000)	freqency	relative frequency
5.0 → 5.15	2	$\frac{2}{28}$
5.15 → 5.30	4	$\frac{4}{28}$
5.30 → 5.45	7	$\frac{7}{28}$
5.45 → 5.60	7	$\frac{7}{28}$
5.60 → 5.75	5	$\frac{5}{28}$
5.75 → 5.90	3	$\frac{3}{28}$
total	28	1

(6d). Both the histogram and the stem-and-leaf plot show that the distribution is mounded -shaped and reasonably symmetric, with no obvious outliers

(6e). Both set of descriptive measures are suitable as the distribution is fairly symmetric with median =mean. The mean and standard deviation would, however, be preferred as they are much easier to use for statistical inference

(6f). Empirical rule does hold since distribution is mound-shaped

(7a). Assuming this data includes all the New York Mets player for 1993, then it would be a population since they only interest is in New York Mets in 1993

(7c). The advantages of a stem-and-leaf plot are that it is simple to make and keeps the actual values of the observations. Here it is very difficult to make a suitable choice for the stems because of the great range of the observations. Also, to get single leaves the observations would have to be rounded to the nearest hundred thus losing the actual observation.

(7e). This distribution is very skewed to the right with 2 likely outliers in the last classes: 5917, 6200

(7f). The mean and standard deviation would not be suitable measures for this distribution, as it is extremely right-skewed with 2 possible outliers on the right. The mean would be pulled to the right by the large values. The median and the interquartiles range would be the appropriate measures to use as they are not affected by extreme observations

(7g). The empirical rule should not be used as this distribution is not mounded-shaped and empirical rule applies to mounded-shaped distribution.