Statistics I, Fall 2012
Homework 02

Ling-Chieh Kung
Department of Information Management
National Taiwan University

1. (50 points) The MS Excel file "StatFa12_hw02.xlsx" contains data from 200 household around the US. Columns A, B, and C contain the amounts spent on food in a year, the amount earned in a year, and the amount of non-mortgage debts. Column D records the region that each household belongs to, where 1 is for northeast, 2 is for midwest, 3 is for south, and 4 is for west. Column E records the location of the household, where 1 means inside a metropolitan area and 2 means the opposite. Answer the following questions based on the data.

   (a) (10 points) Prepare a frequency distribution for the data of annual household income (in Column B). Let the lower bound of the first class be \$20000. Use an equal class interval of \$10000 for all classes. Then draw the histogram.

   (b) (5 points) According to your histogram, is this set of data unimodal, bimodal, or multimodal?

   (c) (5 points) Draw a pie chart to visualize the four proportions of data points that are collected from the four regions.

   (d) (5 points) For each region, find the maximum annual food spending (in Column A), i.e., find the household that spend the most in food and then report its spending. Draw a bar chart to visualize your result.

   (e) (5 points) Recall that we introduced bar charts and pie charts together in class. While you are asked to draw a bar chart for Part (d), do you think a pie chart is a reasonable depiction of the result in Part (d)? Why or why not?

   (f) (10 points) Suppose your boss wants to know how regions and locations affect the average annual food spending. For each of the eight region-location pair, find the average annual food spending for those households in that pair. Present the result in a contingency table. Then try to draw a graph to show your boss the following: "For each region, households inside the metropolitan area in average spend more than those outside the metropolitan area."

   Note. When you go to work, your boss will not tell you which type of chart to draw. It is you that should decide how to present your results (in this case, the eight numbers). Try to find a good way of depiction so that your boss can get the idea in one second. Your grades in this problem depends on your design.

   (g) (5 points) Suppose your boss wants to know whether the income level affects the amount a household spends on food. Draw a graph to suggest your boss an answer. If there is any relationship, describe it. Briefly explain whether the relationship makes sense.

   (h) (5 points) Suppose your boss wants to know whether the non-mortgage debt level affects the amount a household spends on food. Draw a graph to suggest your boss an answer. If there is any relationship, describe it. Briefly explain whether the relationship makes sense.

2. (15 points) Please use the data of non-mortgage debts in the file "StatFa12_hw02.xlsx" to answer the following questions.

   (a) (5 points) Find the mode, mean, variance, and standard deviation. Be careful in determining whether these data form a population or a sample. You may use MS Excel functions if you want and you know which one to use.

   (b) (5 points) Find the first quartile Q_1 and the third quartile Q_3. The find the interquartile range. In calculating the quartiles, apply the formula taught in class. DO NOT use the MS Excel functions QUARTILE() or PERCENTILE() as they are defined in a different way.

   (c) (5 points) Verify Chebyshev's theorem for $k = 1.5$ and $k = 2$, where k is the number of standard deviations from the mean.

3. (10 points) By completing the following table, find the mean absolute deviation, variance, and standard deviation for the following six numbers: 6, 10, 12, 15, 19, and 22. You are encouraged to use software in doing the calculation. Your answer must contain the three measurements AND the completed table.

x_i	$x_i - \mu$	$\lvert x_i - \mu \rvert$	$(x_i - \mu)^2$
6			
10			
12			
15			
19			
22			
Average			

4. (15 points) As Calculus will be required in this semester, let's do some Calculus exercises:

   (a) (5 points) Recall that the population variance is defined as

   $$\sigma^2 \equiv \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}.$$

   Let $j \in \{1, 2, ..., N\}$, find the first-order derivative of σ^2 with respect to x_j.

   (b) (5 points) Find the second-order derivative of σ^2 with respect to μ.

   (c) (5 points) Recall that the population standard deviation is defined as

   $$\sigma \equiv \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}}.$$

   Let $j \in \{1, 2, ..., N\}$, find the first-order derivative of σ with respect to x_j.

5. (10 points) Prove the following way of calculating population means:

   $$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2.$$

   Hint. Show

   $$\sum_{i=1}^{N} (x_i - \mu)^2 = \sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}.$$

   Note. Arithmetic is certainly not the heart of Statistics. However, in order to understand Statistics, some Mathematics is required. Doing these exercises is a good starting point for you to "warm up" for dealing with other mathematical tasks you will encounter in this semester.