Statistics I, Fall 2012
Homework 03

Ling-Chieh Kung
Department of Information Management
National Taiwan University

1. (15 points; 3 points each) In this problem, please answer the following questions based on the frequency distribution for Problem 1a in Homework 2. Assume that you only have grouped data and DO NOT have the ungrouped data.
   (a) Draw an ogive.
   (b) Find the mean and variance for the grouped data by completing the following table:

Class	Frequency	Class midpoint M_i	$(M_i - \bar{x})^2$
$[20, 30)$	8		
$[30, 40)$	26		
$[40, 50)$	31		
$[50, 60)$	67		
$[60, 70)$	32		
$[70, 80)$	28		
$[80, 90)$	6		
$[90, 100)$	2		
Weighted average		$\bar{x} = ?$	$s^2 = ?$

   Note that $\bar{x}$ and s^2 are those for the grouped data.
   (c) Find the mode and standard deviation for the grouped data.
   (d) Find the median for the grouped data.
   (e) As suggested by the mean, median, and mode, are the grouped data symmetric, skewed to the left, or skewed to the right? Briefly explain why.

2. (10 points; 5 points each) In this problem, please use the MS Excel file "StatFa12_hw02.xlsx" used in Homework 2.
   (a) Check whether the empirical rule provides good approximations for one, two, and three standard deviations for the data of non-mortgage debts.
   (b) Are the approximations in Part (a) good or bad? No matter they are good or bad, why? You probably need to draw a graph.

3. (10 points) Given a set of population data $\{(x_i, y_i)\}_{i=1,...,10}$ as below, find the covariance σ_{xy} and correlation coefficient ρ by completing the following table. Then make a conclusion on the relationship between the two variables.

i	x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)(y_i - \mu_y)$
1	7	5			
2	4	6			
3	2	9			
4	12	6			
5	10	15			
6	7	6			
7	8	9			
8	8	15			
9	6	9			
10	3	3			
Average	$\mu_x = ?$	$\mu_y = ?$	–	–	$\sigma_{xy} = ?$

4. (5 points) Prove that the correlation coefficient is always within -1 and 1.

   Hint. Use the Cauchy-Schwarz inequality.

   Note. I certainly know it is very easy to find this proof online. All I hope is that when you saw $-1 \leq \rho \leq 1$ in class, there was one second that you wondered why. I hope you are curious about this and not satisfied by the intuition we discussed in class. Try to construct (or at least find) the proof and try to write it down carefully. Clearly define variables and notations you use and make the derivations and logic sound. We will not be strict in grading proofs. Remember, you may copy anything you find by yourself EXCLUDING your friend's proof!

5. (12 points; 4 points each) For a given set of two-dimensional data $\{(x_i, y_i)\}_{i=1,\ldots,N}$, suppose $y_i = a + bx_i$ for all $i = 1, 2, ..., N$ for some known constants a and b. Let μ_x and σ_x^2 be the mean and variance for x_is, respectively.

   (a) Prove that the mean for y_is is $\mu_y = a + b\mu_x$.

   (b) Prove that the variance for y_is is $\sigma_y^2 = b^2\sigma_x^2$.

   (c) Given the above two results, I tried to prove that ρ, the correlation coefficient of x_is and y_is, is always 1. Consider my proof below:

   Proof. Let σ_{xy} be the covariance of x_is and y_is. By applying the facts that $y_i = a + bx_i$ and $\mu_y = a + b\mu_x$, we have

$$\begin{aligned} \sigma_{xy} \equiv \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N} &= \frac{\sum_{i=1}^{N}(x_i - \mu_x)(a + bx_i - a - b\mu_x)}{N} \\ = \frac{b\sum_{i=1}^{N}(x_i - \mu_x)^2}{N} &= b\sigma_x^2. \end{aligned}$$

   Now, by applying the fact that $\sigma_y^2 = b^2\sigma_x^2$, we have

$$\rho \equiv \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{b\sigma_x^2}{\sigma_x(b\sigma_x)} = 1.$$

   Therefore, the correlation coefficient $\rho = 1$. □

   Determine whether the proof is correct. If it is, just write down "Yes." Otherwise, clearly indicate where the error(s) is and then fix the error(s).

   Hint. For the first two questions, the key is to follow the definitions of mean and variance. Clearly write down the definitions and then apply the fact $y_i = a + bx_i$.

6. (10 points; 2 points each; MODIFIED from Problem 4.2 in the textbook) Given three sets $A = \{1, 3, 5, 7, 8, 9\}$, $B = \{2, 4, 7, 9\}$, and $C = \{1, 2, 3, 4, 6\}$, please solve for the following:

   (a) $A \cup C$.

   (b) $A \cap B$.

   (c) $A \cap B \cap C$.

   (d) $(A \cup B) \cap C$.

   (e) $(B \cap C) \cup (A \cap B)$.

7. (11 points; MODIFIED from Problem 4.36 in the textbook) Consider the following cross tabulation in which the frequencies of twelve joint events are recorded.

	D	E	F	G
A	3	9	7	12
B	8	4	8	4
C	9	5	3	7

For example, event $A \cap D$ has occurred for three times. Now, calculate the following probabilities:

(a) (1 point) $\Pr(A)$.

(b) (1 point) $\Pr(A \cap F)$.

(c) (2 points) $\Pr(A|F)$.

(d) (2 points) $\Pr(B \cup E)$.

(e) (3 points) $\Pr(D \cup G|C)$.
Hint. Among all the occurrences of C, for how many times does the join event $D \cup G$ occur?

(f) (2 points) Are the two experiments independent? Why or why not?

8. (15 points; 3 points each) In a department, undergraduate students can be categorized by the years they enter the university. In particular, there are four classes of students: freshmen, sophomores, juniors, and seniors. In the freshmen class, there are 10 girls and 40 boys. In the sophomore class, there are 15 girls and 35 boys. In the junior class, there are 12 girls and 38 boys. In the senior class, there are 18 girls and 32 boys.

   (a) Use the class of a student as one variable and the sex as the other variable to construct a joint probability table.

   (b) Among all students in the department, what is the proportion of girls? Note that this proportion is exactly the probability that a student randomly selected from the whole department is a girl.

   (c) Among all students in the sophomore class, what is the proportion of girls? Note that this proportion is exactly the probability that a student randomly selected from the sophomore class is a girl.

   (d) Is the probability in Part (b) a marginal, joint, union, or conditional probability? How about the probability in Part (c)?

   (e) Based on your answers in Parts (c) and (d), are the class and sex of a student dependent or independent? Briefly explain why, in words or using equations.

9. (12 points; 4 points each; MODIFIED from Problem 4.18 in the textbook) According to the nonprofit group Zero Population Growth, 78% of US adults live in urban areas. Scientists at Princeton University and the University of Wisconsin report that about 18% of all US adults are taking care of their ill relatives. Suppose that among all US adults living in urban areas, 13% are taking care of their ill relatives.

   (a) Determine the probability that a randomly selected US adult is living in an urban area and taking care of her ill relatives.

   (b) Construct a complete joint probability table for two variables: Whether one lives in urban areas (outcomes: yes and no) and whether one is taking care of her ill relatives (outcomes: yes and no).

   (c) Determine the conditional probability of an US adult living in a nonurban area given that she is taking care of her ill relatives.