Statistics and Data Analysis, Fall 2015
Midterm Exam

Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

Note. This exam is in-class and open everything (including all kinds of electronic devices). However, an exam taker is not allowed to communicate with any person during the exam. Cheating will result in severe penalty. All data needed for this exam is contained in the MS Excel file "SDA-Fa15_midterm_data.xlsx." You do not need to return the problem sheet. The maximum number of points that one may earn is 100.

1. (20 points) Consider a set of population data with 1000 values contained in the sheet "Frequency."

   (a) (3 points) Construct a frequency distribution with classes $[0, 100)$, $[100, 200)$, ..., and $[600, 700)$.

   (b) (3 points) Draw a histogram with classes $[0, 100)$, $[100, 200)$, ..., and $[600, 700)$.

   (c) (4 points) Find the mean, median, and mode. Is this set of data skewed to the right, skewed to the left, or symmetric? How to determine the skewness by comparing the mean, median, and mode?

   (d) (3 points) Find the variance, standard deviation, and coefficient of variation.

   (e) (3 points) Let μ and σ be the mean and standard deviation. How many values are above $\mu + 3\sigma$ or below $\mu - 3\sigma$? If we remove these values, what will be the new value of the standard deviation? Intuitively why it becomes smaller (or bigger, according to your observation).

   (f) (4 points) Ignore Part (e) and consider the 1000 values. Find the mean absolute deviation. Is it less or greater than the standard deviation? Intuitively explain why.

2. (15 points) Consider a set of sample data with 49 rows of values contained in the sheet "Grades." These are the midterm grades of 49 students in Accounting, Statistics, and Economics.

   (a) (3 points) By comparing their standard deviations, which subject has the largest variability?

   (b) (2 points) Draw a scatter plot for the first ten students' Accounting and Economics grades (the data are contained in rows 2 to 11).

   (c) (5 points) Find the three correlation coefficients, one for each pair of subjects. Interpret these correlation coefficients.

   (d) (5 points) A way to determine the relationship between two variables is through a *contingency table*. In this problem, let's construct a two-by-two contingency table for Accounting and Statistics. First, calculate the means for these two variables. Second, for each student, determine whether she/he is above or below average in Accounting and Statistics. Finally, find the numbers of students that should be put into the following table. E.g., (1) is the number of students who are below average in both subjects.

Accounting	Statistics	
	Below average	Above average
Below average	(1)	(2)
Above average	(3)	(4)

   Complete the table. According to the information in the table, intuitively comment on the relationship between Accounting and Statistics grades. Obviously, your answer should share some common idea with that from the correlation coefficient.

3. (10 points) Consider an unfair coin whose probability of getting a head is 0.4. Let X_n be the number of heads obtained after tossing the coin for n times.

   (a) (3 points) If $n = 5$, find the mean and standard deviation of X_n.

(b) (3 points) If $n = 50$, find $\Pr(X_n \leq 25)$.

(c) (4 points) If $n = 10$, find $\Pr(X_n = 0)$.

4. (15 points) Consider a dice whose probability of getting each value is $\frac{1}{6}$. Let X be the numbers of dice rolling until we see 5 or 6 as an outcome. For example, suppose that we roll the dice and get 1, 4, 1, 2, 2, 3, 5 in order, then $X = 7$ in this trial. The sample space of X is all positive integers. Obviously, $\Pr(X = 1) = \frac{1}{3}$.

   (a) (3 points) Find $\Pr(X = 2)$.

   (b) (3 points) Find $\Pr(X = 3)$.

   (c) (4 points) Find $\Pr(X = k)$ for any nonnegative integer k.

   (d) (5 points) Try your best to find the expected value of X. You may rigorously derive the answer. If you cannot do this, you may numerically find $\Pr(X = k)$ for $k = 1, 2, ..., n$, where n is "large enough," numerically calculate the expected value, and finally make your estimation. If you cannot do either, you may guess one number and intuitively explain why. Points will be given to you according to the correctness and persuasiveness of your answer.

5. (15 points) Let's fit a theoretical distribution to a set of observed sample data.

   (a) (5 points) Consider the data contained in the sheet "Frequency." Find the observed frequencies of classes $[0, 100)$, $[100, 200)$, ..., and $[600, 700)$. Then let X be uniform between 0 and 700 and find the theoretical frequencies of X according to the same set of classes. Write down all the observed and theoretical frequencies. By comparing them, comment on whether it is appropriate to say that these 1000 values follow a uniform distribution between 0 and 700.

   (b) (10 points) Consider the Statistics grades contained in the sheet "Grades." Find the observed frequencies of classes $[40, 50)$, $[50, 60)$, ..., and $[90, 100)$. Then let $X \sim \text{ND}(70, 10)$ and find the theoretical frequencies of X according to the same set of classes. Write down all the observed and theoretical frequencies. By comparing them, comment on whether it is appropriate to say that Statistics grades follow a normal distribution with mean 70 and standard deviation 10.

6. (25 points) Let X_i be the outcome of rolling a fair dice for the ith time. Let $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$ be the sample mean of rolling a fair dice n times.

   (a) (3 points) Find the mean and standard deviation of X_i.

   (b) (3 points) Find the probability distribution of X_i.

   (c) (3 points) Find the mean and standard deviation of $\bar{x}_2$.

   (d) (5 points) Find the probability distribution of $\bar{x}_2$.

   (e) (3 points) Find the probability distribution of $\bar{x}_{100}$.

   (f) (3 points) Find $\Pr(\bar{x}_{100} > 3.6)$.

   (g) (5 points) Suppose that you roll a dice 100 times and observe the sample mean as 2.5. Comment on this observation and its implication on whether the dice is fair or not.

7. (10 points; 5 points each) Please answer True or False. DO NOT provide any reason.

   (a) The citizenship of a person is of nominal scale.

   (b) When the sample size increases, the sample mean will become more volatile.

   (c) If two variables have zero correlation coefficient, they are independent.

   (d) For a given sample, the sample variance and the variance of the sample mean are different.

   (e) If the population is normally distributed, the sample mean will also be normally distributed. The sample size does not matter.