STAT 2507 and BIT 2000 A
SOLUTIONS TO MIDTERM TEST (Version 1)
Fall 2014

1. Consider the following stem-and-leaf plot of 23 observations.

   Leaf unit=1

   0 | 012

   1 | 2248

   2 | 1134599

   3 | 00036

   4 | 457

   5 | 1

   The upper quartile, Q_3, and the median are, respectively:

   (a) 31.5 and 24.5 **(b)** 36 and 24 **(c)** 30.75 and 25 **(d)** 33 and 25

   Solution: The position of the median is $0.5(n+1) = 0.5(24) = 12$, and the 12th element in the ordered data set is 25. The position of the upper quartile is $0.75(n+1) = 0.75(24) = 18$, and the 18th element in the ordered data set is 33.

2. Suppose you have a data set that contains a small fraction of the observations that are very large or very small. Which of the following numerical measures would be most appropriate for describing variation in the data set?

   (a) Sample standard deviation **(b)** Sample range
   (c) Sample interquartile range **(d)** All of the above

   Solution: The sample range and the sample standard deviation are too much sensitive to the presence of outliers. At the same time, the sample interquartile range is robust to the presence of outliers. Therefore, when the data set contains outliers, the sample interquartile range would be most appropriate as a measure of spread or variation.

3. Which of the following randomly selected measurements, X, may be considered to be an outlier if it was selected from the given population?

   (a) $X = 10$ from a population with $\mu = 3$ and $\sigma^2 = 16$.
   (b) $X = 1/2$ from a population with $\mu = -3$ and $\sigma^2 = 1$.
   (c) $X = -1$ from a population with $\mu = 1$ and $\sigma^2 = 25$.
   (d) $X = 0$ from a population with $\mu = -2$ and $\sigma^2 = 3$.

   Solution: We know that observations with z-scores equal to or greater than 3 are considered very unlikely. The only observation among (a)–(d) that has the z-score in excess of 3 is $X = 1/2$.

4. If the interval $\bar{x} \pm s$ is equal to (3,7) then the interval $\bar{x} \pm 2s$ is

   (a) (1,9) **(b)** (2,5) **(c)** (2,8) **(d)** Impossible to compute

   Solution: By the statement of the problem, $\bar{x} - s = 3$ and $\bar{x} + s = 7$. Thus we have two linear equations with two unknowns. Solving this equations yields $\bar{x} = 5$ and $s = 2$. Therefore, $\bar{x} - 2s = 5 - 2(2) = 1$ and $\bar{x} + 2s = 5 + 2(2) = 9$.

5. Which of the following statements is/are always true ?

   I. If a quantitative variable has a finite number of possible values to take, it is discrete.
   II. If a quantitative variable has an infinite number of possible values to take, it is continuous.

   (a) I only **(b)** II only **(c)** Both **(d)** Neither

   Solution: Both discrete and continuous variables can take on an infinite number of possible values. At the same time, the set of values of a continuous random variable is always infinite. Therefore the correct answer is (a).

6. It is said that the normal range of systolic blood pressure is 90–130 mmHg. From a random sample of 1,000 high school students, sample mean and sample standard deviation of the systolic blood pressure were found to be $\bar{x} = 110$ mmHg and $s = 10$ mmHg. Which of the following would always be true about the sample?

   (a) Approximately 750 students have their systolic blood pressure in the normal range.
   (b) At most 750 students have their systolic blood pressure in the normal range.
   (c) At least 750 students have their systolic blood pressure in the normal range.
   (d) No statement can be made on this sample of students.

   Solution: According to Chebyshev's inequality, at least 75% of sample data belong to the internal $\bar{x} \pm 2s$. In this problem, $(\bar{x} - 2s, \bar{x} + 2s) = (90, 130)$. Therefore the correct answer is (c).

7. Suppose the sample correlation coefficient r between X and Y in a bivariate data set is equal to $r = 0.85$. If the sample variance of X is greater than that of Y, how would you describe the slope of the least squares regression line of Y on X?

   (a) The estimated slope is the same as 0.85.
   (b) The estimated slope is positive but less than 0.85.
   (c) The estimated slope is greater than 0.85.
   (d) The estimated slope is not related to the correlation coefficient.

   Solution: We know that the slope b of the least-squares line satisfies $b = r\dfrac{s_y}{s_x}$. Since $0 < s_y = \sqrt{s_y^2} < \sqrt{s_x^2} = s_x$, it follows that $0 < b < r = 0.85$.

8. When you describe or summarize data, certain degree of information may be lost. Which of the following diagrams or measures preserves the largest amount of information?

   (a) Histogram **(b)** Boxplot **(c)** Median **(d)** Stem-and-leaf plot

   Solution: A boxplot is constructed by using five summary measures (and outliers) and ignoring the rest of the data, and the median is just one of the five summary measures used

to construct the boxplot. With a histogram, the loss of information occurs due to grouping, whereas no information is lost with a stem-and-leaf plot because you can see every data point.

9. Event A occurs with probability 0.4 and event B occurs with probability 0.7. If we know that the event A has occurred, the probability that B will occur is 0.7. If we know that event B has occurred, then what is the probability that event A will occur, i.e., what is $P(A|B)$?

(a) 0.60 **(b)** 0.28 **(c)** 0.70 **(d)** 0.40

Solution: Using the Multiplication Rule for conditional probability, we obtain

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{(0.7)(0.4)}{(0.7)} = 0.4.$$

10. Dan walks to work 20% of days and the rest of the days he takes the bus to work. 40% of the days that he walks to work he arrives late while he arrives late only 5% of the days he takes the bus. What is the probability that Dan arrives late in a day?

(a) 0.800 **(b)** 0.120 **(c)** 0.128 **(d)** 0.080

Solution: Let W and B be the events that Dan walks to work today and takes a bus to work today, respectively. Let L be the event that Dan arrives late today. By assumption

$$P(W) = 0.2, \quad P(B) = 0.8, \quad P(L|W) = 0.4, \quad P(L|B) = 0.05.$$

We need to find $P(L)$. Using the Law of Total Probability, we get

$$P(L) = P(L|W)P(W) + P(L|B)P(B) = (0.4)(0.2) + (0.05)(0.8) = 0.12.$$

11. Under the conditions of problem 10, if we know that Dan was late today, what is the probability that he walked to work?

(a) 2/3 **(b)** 1/3 **(c)** 2/12 **(d)** 1/12

Solution: Using the notation and result of problem 10 and applying Bayes' Theorem, we get that the required probability is equal to

$$P(W|L) = \frac{P(L|W)P(W)}{P(L)} = \frac{(0.4)(0.2)}{0.12} = \frac{0.08}{0.12} = \frac{2}{3}.$$

12. A Professor gives her students 8 questions a week before an exam and announces that 5 of these questions will be chosen for the exam. If Sarah, a student in the class, knows the solutions to 6 of the questions, then what is the probability that Sarah will receive a perfect mark on her exam next week?

(a) 5/6 **(b)** 6/8 **(c)** 5/56 **(d)** 6/56

Solution: Using the classical definition of probability,

$$P(\text{Sarah will receive a perfect mark}) = \frac{\text{number of favourable outcomes}}{\text{number of possible outcomes}} = \frac{C_5^6}{C_5^8} = \frac{\frac{6!}{5!1!}}{\frac{8!}{5!3!}} = \frac{6}{56}.$$

13. An interior decorator must furnish two offices that already have a desk and a chair. One office needs 1 file cabinet and 1 bookcase and the other needs 1 file cabinet and 2 bookcases. At a local office furniture store, there are 5 models of file cabinets, and 6 models of bookcases all of which are compatible. How many choices does the decorator have if he wants to select 2 file cabinets and 3 bookcases but he does not want to select more than one of any model?

(a) 30 **(b)** 180 **(c)** 150 **(d)** 200

Solution: There are C_2^5 different ways in which 2 models of file cabinets can be selected by the decorator from the 5 models available, and, regardless of which particular 2 models of file cabinets have been chosen by the decorator, there are C_3^6 different ways in which 3 models of bookcases can be selected by the decorator from the 6 models available. Therefore, by the Multiplication Rule, the total number of choices that the decorator can have is $C_2^5 \times C_3^6 = 200$.

14. A probability distribution of a discrete random variable X is partially given in the following table, with the additional information that $p(1) = 3p(5)$. Determine the missing entries in the table.

x	0	1	2	3	4	5
$p(x)$	0.20	?	0.25	0.15	0.20	?

(a) $p(1) = 0.30, \quad p(5) = 0.10$
(b) $p(1) = 0.15, \quad p(5) = 0.05$
(c) $p(1) = 0.05, \quad p(5) = 0.15$
(d) $p(1) = 0.30, \quad p(5) = 0.10$

Solution: By the normalization property of a pmf,

$$p(0) + p(1) + p(2) + p(3) + p(4) + p(5) = 1.$$

Therefore, using the fact that $p(1) = 3p(5)$, we obtain

$$0.20 + 3p(5) + 0.25 + 0.15 + 0.20 + p(5) = 1.$$

This gives $4p(5) = 0.20$, and hence $p(5) = 0.05$ and $p(1) = 3p(5) = 0.15$.

15. In a certain part of downtown Ottawa, a car that is illegally parked on a street will be fined \$30 if caught, and the chance of being caught is 60%. What is the expected fine for person who parked on this street?

(a) \$12.00 **(b)** \$10.00 **(c)** \$18.00 **(d)** \$15.00

Solution: Let X be the cost of fine. Then X takes on only two values: 30 with probability 0.6 and 0 with probability 0.4. Therefore $E(X) = 30(0.6) + 0(0.4) = 18$.

16. When the price of gasoline gets high, consumers become very concerned about the gas mileage obtained by their cars. One consumer was interested in the relationship between car engine size (number of cylinders) and gas mileage (litres per 100 km). The consumer took a random sample of 7 cars and recorded the following information:

$$\sum_{i=1}^{7} x_i = 40, \quad \sum_{i=1}^{7} y_i = 77, \quad \sum_{i=1}^{7} x_i y_i = 488, \quad s_x = 1.799, \quad s_y = 4.583.$$

Fit the least-squares line relating car engine size, x, and fuel efficiency, y, and find the predicted fuel efficiency for a car with a 6-cylinder engine. Round all intermediate numbers using 3 decimal places.

(a) 16.25 litres per 100 km
(b) 9.35 litres per 100 km
(c) 11.71 litres per 100 km
(d) 20.12 litres per 100 km

Solution: The least-squares line is given by the formula

$$y = a + bx, \quad \text{where} \quad b = \frac{s_{xy}}{s_x^2} \quad a = \bar{y} - b\bar{x},$$

and

$$s_{xy} = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{(n-1)} = \frac{488 - (40)(77)/7}{7-1} = 8,$$

$$b = \frac{s_{xy}}{s_x^2} = \frac{8}{(1.799)^2} = \frac{8}{3.236} = 2.472,$$

$$a = \bar{y} - b\bar{x} = (77)/7 - (2.472)(40)/7 = -3.125.$$

Therefore the least-squares line is given by

$$y = -3.125 + 2.472x,$$

and the predicted value of fuel efficiency for a car with a 6-cylinder engine is

$$-3.125 + 2.472(6) = 11.707 \approx 11.71 \text{ litres per km .}$$