40.004 Statistics 2018: **Problem set 1**

due 13 February, 2018 at 11:59 pm. Submit on ***e-dimension***.

1. A radio station in Singapore frequently broadcasts an opinion question during the morning rush hour and gives a telephone number to sms/text "Yes" or "No" responses. Poll results are declared at the end of the morning rush hour. Are such call-in polls more likely to be fair or more likely to be biased?

2. To determine whether the artificial sweetener aspartame causes headaches, researchers gave capsules containing aspartame or placebo to subjects and observed their responses. After a period which allowed the subjects to rid their bodies of chemicals, those originally given placebo were given aspartame and vice versa. Similar rates of headaches were reported for both groups.

   (a) What are the control and treatment groups in this study?

   (b) This is an example of a *cross-over* design, in which each person is assigned both treatments in random order. What advantages does this design have over a study in which people receive only one treatment?

   (c) If the study noticed, among other things, that one of the treatment groups suffered higher than average rates of stomach upsets, would the researchers be justified in concluding that aspartame can cause stomach upsets?

3. A school district plans to survey 1000 out of 50,000 parents with enrolled children regarding their preferences on enrolment. A complete alphabetical list of parent names is available. In each of the following, name the sampling method used.

   (a) One of the first 50 names on the complete list is randomly chosen; that person and every 50th person on the list after that person are surveyed.

   (b) The complete list is divided into separate lists by their oldest child's year in school. Random numbers are assigned to names, and each list is ordered by the assigned random numbers. The first 2% of the ordered names in each list are surveyed.

   (c) A random number is assigned to each name using a random number generator. Names are ordered by the random numbers and the first 1000 are surveyed.

4. The following are the pH measurements on 50 soil samples a landscaper took from ground adjacent to a new building.

6.10	6.74	6.22	5.65	6.38	6.70	7.00	6.43	7.00	6.70	6.70	5.94	6.28
6.34	6.62	6.55	2.92	6.10	6.20	6.70	7.00	6.85	6.31	6.26	6.36	6.28
6.38	6.70	6.62	7.00	6.45	6.31	2.86	6.31	6.09	6.17	6.64	6.45	
7.00	6.18	6.58	5.38	6.34	7.00	5.70	6.65	6.56	6.00	6.70	6.45	

   (a) Calculate the five number summary statistic of these data. Does this suggest a symmetric or a skewed distribution.

   (b) Compute a 10% or 1/10-trimmed mean and compare it with the sample mean? Does this comparison suggest outliers?

   (c) Compute the IQR and standard deviation of the sample.

   (d) Make a box plot. Are there outliers (using the 1.5 IQR rule)?

   (e) Create a normal probability plot. Does this data look normal? Does this look skewed?

5. The *Land Transport Authority* in Singapore wishes to conduct a survey to determine the proportion of free-riders on Singapore buses, given by π (unknown). A sample of size n is chosen for survey. Each participant is asked to toss a coin whose probability of showing a head is q. If the coin shows a head, the person is asked to answer Question A; otherwise they answer Question B.

   - Question A: Have you free-ridden in a bus in Singapore in 2017?
   - Question B: Were you born in February?

   Suppose proportion p of the responses are 'Yes'.

   (a) Estimate π; state any assumptions.
   (b) Is it possible for your estimate to lie outside the range $[0, 1]$?
   (c) Are there any unhelpful values of q?

6. Hospitals are graded based on their success rates in treating different categories of patients. We want to compare two hospitals - A, which is a university-affiliated research hospital, and B, which is a general community hospital – with respect to success rates for a certain complicated surgery. The data classified by low risk patients and high risk patients are shown in the following table.

	Low Risk				High Risk		
	Success	Failure	Total		Success	Failure	Total
Hospital A	400	100	500	Hospital A	160	640	800
Hospital B	300	200	500	Hospital B	20	180	200
Total	700	300	1000	Total	180	820	1000

   (a) Calculate the success rates for each category of patients for both hospitals. Which hospital is better?
   (b) Aggregate the data over the two categories of patients and calculate the overall success rates for both hospitals. Now which hospital is better?
   (c) Explain the discrepancy between the results obtained in (a) and (b).

7. The monthly average unleaded gasoline prices per gallon, for the 55 largest cities in the U.S., from Jan 1996 to Dec 1997 are given below.

   112.9 112.4 116.2 125.1 132.3 129.9 127.2 124.0 123.4 122.7 125.0 126.0
   126.1 125.5 123.5 123.1 122.6 122.9 120.5 125.3 127.7 124.2 121.3 117.7

   (a) Make a time-series plot of the data.
   (b) Predict the gasoline price for January 1998 by computing the EWMA (and the associated MAPE) using 3 different parameters: $\alpha = 0.3, 0.4, 0.5$. Which α is the best?

8. A *population* consists of $N = 4$ numbers: 1, 2, 3, 4. Each number has equal chance of being selected.

   (a) Calculate the population variance σ^2.
   (b) A random sample of size $n = 2$ is selected *with* replacement from the population. Find the sampling distribution for the sample variance (e. g. find all the possible values of the sample variance and the corresponding frequencies).
   (c) Calculate the expected value of the sample variance from (b), and compare it with the population variance obtained in (a).

9. A random sample $X_1, \ldots, X_{150}$ from a population with mean $\mu = 40$ and standard deviation $\sigma = 15$ but an unknown distribution. Let

$$U = \frac{X_1 + \ldots + X_{50}}{50}, \qquad V = \frac{X_{51} + \ldots + X_{150}}{100}.$$

   (a) What are the approximate distributions of U and V?
   (b) Which probability would you expect to be larger, $\Pr(35 \leq U \leq 45)$ or $\Pr(35 \leq V \leq 45)$? Why?
   (c) Find $\Pr(35 \leq U \leq 45)$ and $\Pr(35 \leq V \leq 45)$ using the normal approximation.