40.004 Statistics 2018: **Problem set 1 solutions**

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?

   Solution: The poll is conducted during morning rush hours, when most listeners are people with regular working hour jobs, creating a convenience sample of employed, more affluent adults. Hence the survey only samples this particular segment of the population and induces bias in the survey results. Moreover, in "call-in" polls, people who are strongly for or against an issue are likely to voice their opinions. The results hence often reflect proportions of people with strong opinions which may differ from the overall proportions for or against an issue.

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.

   Solution:

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

   Control group: subjects taking placebo. Treatment group: subjects given aspartame.

   (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?*

   A cross-over design reduces unpredictable variability. For instance, even with randomization, it may be the case that one group ends up with more people prone to headaches (due to stress level, health, etc). A cross-over design helps to cancel out such factors; each person is effectively serving as his or her own control.

   (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?*

   No, the research was not aimed at investigating stomach upsets, and the higher incidence may be simply due to chance. To properly investigate the effects of aspartame on stomach upsets, a new experiment needs to be carried out.

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.

   Solution:

   (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.*

   Systematic sampling.

   (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.*

   Stratified sampling (the strata are defined by oldest child's year at school).

   (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.*

   Simple random sampling.

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	

Solution: Let the data be denoted $x_1, x_2, \ldots, x_{50}$ with its order statistics given by $x_{(1)} \leq x_{(2)} \leq \ldots x_{(50)}$.

(a) *Calculate the five number summary statistic of these data. Does this suggest a symmetric or a skewed distribution.*
Five number summary statistics from R:

$$\text{Min} = 2.86, Q_1 = 6.205, \text{Median} = 6.405, Q_3 = 6.70, \text{Max} = 7.00.$$

The formula from book gives $Q_1 = 6.195$. Note the Q_1 and Q_3 are approximately equidistant from the median, hence the summary suggests that the distribution is symmetric.

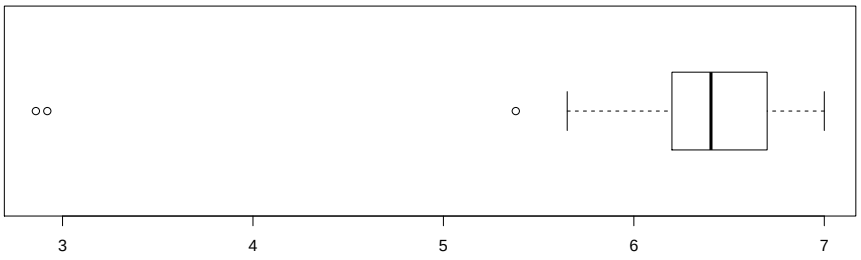
(b) *Compute a 10% or 1/10-trimmed mean and compare it with the sample mean? Does this comparison suggest outliers?*
The mean of the data set is 6.297 and the 10% trimmed mean is:

$$T_{1/10} := \frac{x_{(6)} + \ldots + x_{(45)}}{50 - 2 \times 5} = 6.434.$$

So $T_{1/10}$ is a bit higher than the mean suggesting some outliers on the lower end of the data.

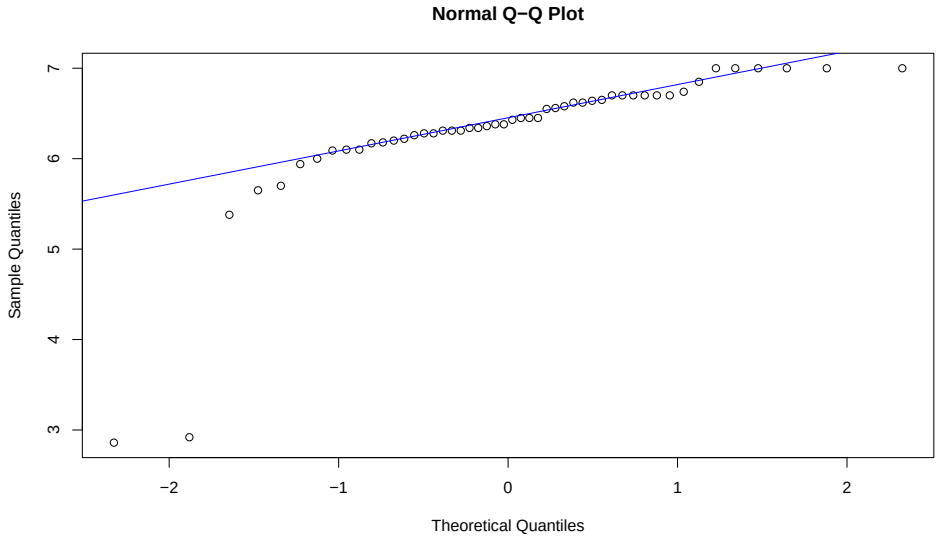
(c) *Compute the IQR and standard deviation of the sample.*
Here IQR = 0.495 (using formula from the book it is 0.505). The standard deviation is $s = 0.788$. Incidentally an approximate calculation of the standard deviation using IQR is $s = \text{IQR}/1.34 = 0.369$ which is quite a bit lower than the actual value suggesting that there may be outliers in the data set.

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



Using the $1.5 \times IQR$ rule there are three points beyond the whiskers of the data suggesting outliers.

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



Other than the outliers the data looks quite close to normal.

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'. (think k responses are 'Yes' and $p = k/n$.)

   Solution:

   (a) *Estimate π; state any assumptions.*
   Assumptions:
   – The surveyed individuals do not lie.
   – Assuming that birthdays are uniformly distributed, the proportion of people born *in February* is about $28/365 \approx 1/13$ (or considering 12 months we may have $1/12$). Either answer is fine.

   Proportion q of the people will answer question A, and only those who free-rode will answer Yes; proportion $(1-q)$ will answer question B, and only those born in February will answer Yes.
   Hence we expect

$$q\,\pi + (1-q)\,\frac{1}{13} = p,$$

   which gives the estimate

$$\hat{\pi} = \frac{p}{q} - \frac{1-q}{13q}.$$

   (b) *Is it possible for your estimate to lie outside the range $[0,1]$?*
   If $p < (1-q)/13$, then the above estimate becomes negative; if $p > (1+12q)/13$, then the estimate becomes greater than 1.

   (c) *Are there any unhelpful values of q?*
   Here $q = 0$ and $q = 1$ are unhelpful, since people may quickly figure out that the coin is biased. Moreover, if $q = 0$, then we cannot actually recover π, since the expression for π has q in the denominator.

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?*

Success rates	Risk	
Hospital	Low	High
A	80%	20%
B	60%	10%

In this regard, Hospital A 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?*

Hospital	Success rate
A	43%
B	46%

Overall, Hospital B has a better success rate.

(c) *Explain the discrepancy between the results obtained in (a) and (b).*

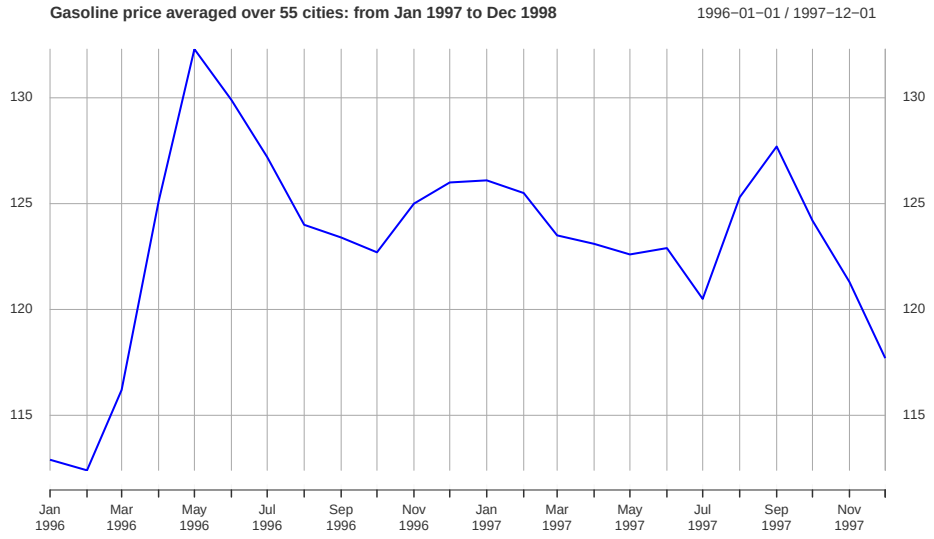
This is a classic example of Simpson's paradox. While Hospital A has a higher success rate for both risk groups, it has a larger percentage of high risk patients than Hospital B. Since the high risk patients have a lower success rate, the discrepancy brings Hospital A's overall success rate below that of Hospital 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

Solution:

(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?*

Let $x_1, \ldots, x_{24}$ be the gasoline prices. We intend to estimate x_{25}. The EWMA estimates (with MAPE) are given by

$$\begin{aligned} \alpha = 0.3: \ \hat{x}_{25} &= \text{EWMA}_{24} = 121.78, \quad \text{MAPE} = 11.67\%, \\ \alpha = 0.4: \ \hat{x}_{25} &= \text{EWMA}_{24} = 121.10, \quad \text{MAPE} = 8.27\%, \\ \alpha = 0.5: \ \hat{x}_{25} &= \text{EWMA}_{24} = 120.40, \quad \text{MAPE} = 6.42\%. \end{aligned}$$

The predictions are calculated recursively using the formula $\text{EWMA}_t = \alpha\, x_t + (1-\alpha)\, \text{EWMA}_{t-1}$. Among these values of α, 0.5 is the best as it gives the least MAPE.

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).

Solution:

(a) The population variance is given by

$$\sigma^2 = \mathbb{E}\left((X-\mu)^2\right) = \frac{1}{4}(1-2.5)^2 + \frac{1}{4}(2-2.5)^2 + \frac{1}{4}(3-2.5)^2 + \frac{1}{4}(4-2.5)^2 = 1.25.$$

(b) There are $4^2 = 16$ random samples in total: $\{1,1\}$, $\{1,2\}$, $\ldots$, $\{4,4\}$.
For $\{1,1\}$, $s^2 = 0$; for $\{1,2\}$, $s^2 = 0.5$, etc. After computing s^2 for all 16 samples, we find the following sampling distribution

s^2	0	0.5	2	4.5
Probability	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{2}{16}$

(c) The expected value, or average, of the sample variance is therefore

$$0 \times \frac{4}{16} + 0.5 \times \frac{6}{16} + 2 \times \frac{4}{16} + 4.5 \times \frac{2}{16} = 1.25,$$

in agreement with part (a). This is expected, since we know that $\mathbb{E}(s^2) = \sigma^2$.

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

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

Solution:

(a) *What are the approximate distributions of U and V?*
U is approximately normal with mean $\mu = 40$ and SD $= \sigma/\sqrt{n} = 15/\sqrt{50} = 2.121$.
V is approximately normal with mean $\mu = 40$ and SD $= \sigma/\sqrt{n} = 15/\sqrt{100} = 1.5$.

(b) *Which probability would you expect to be larger,* $\Pr(35 \le U \le 45)$ *or* $\Pr(35 \le V \le 45)$*? Why?*
Since V has a smaller standard deviation, more of the probability clusters close to the mean of 40, so we expect $\Pr(35 \le V \le 45)$ to be larger than $\Pr(35 \le V \le 45)$.

(c) *Find* $\Pr(35 \le U \le 45)$ *and* $\Pr(35 \le V \le 45)$ *using the normal approximation.*
Using normal approximation we have the following values with $Z \sim \mathcal{N}(0,1)$:

$$\Pr(35 \le U \le 45) = \Pr\left(\frac{35-40}{2.121} \le \frac{U-40}{2.121} \le \frac{45-40}{2.121}\right)$$
$$\Pr(-2.357 \le Z \le 2.357) = \Phi(2.357) - \Phi(-2.357) = 0.9818.$$
$$\Pr(35 \le V \le 45) = \Pr\left(\frac{35-40}{1.5} \le \frac{V-40}{2.121} \le \frac{45-40}{1.5}\right)$$
$$\Pr(-3.333 \le Z \le 3.333) = \Phi(3.333) - \Phi(-3.333) = 0.9992.$$