SUTD ESD

# Statistics 2017

## Assignment 1 Solutions

**Question 1.** There are many problems with the experimental design. For instance, there is no control group, so the participants may be swimming faster simply because they practiced more, as opposed to because they used a better technique. Also, the swimmers were not randomly chosen, but were volunteers, so their commitment and abilities may already be above the group average.

**Question 2.** (a) Systematic.

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

(c) Simple random.

**Question 3.** (a) Control: subjects taking placebo. Treatment: subjects given aspartame.

(b) 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) 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.

**Question 4.** (a) The treatment factor is the techniques involved (let us denote them by $A$, $B$ and $C$). The noise factor is the workers, since they may work at different rates.

(b) There are many ways. One possibility is:
Worker 1: $\{A, B, C, C, A, B\}$
Worker 2: $\{B, C, A, C, B, A\}$
Worker 3: $\{C, C, B, A, B, A\}$
Worker 4: $\{A, C, A, B, B, C\}$
In this design, we have ensured that each worker uses each technique twice, but in random order.

(c) Again, there are many possible solutions. One plausible solution is to pick 12 of those portions, and assign 3 portions to each worker; each worker then uses each technique once, in random order.

Another possibility is to pick 3 out of the 4 workers randomly, then assign 6 portions to each of these workers, and implement the scheme from part (b).

It is also possible to give each worker 5 portions, though this has the disadvantage that not every technique will be tested an equal number of times.

**Question 5.** (a) Mean: 98.563, median: 98.6, standard deviation: 0.751, min: 97.0, max: 99.8, $Q_1$: 98.2 or 98.125 (depends on which method you use), $Q_3$: 99.1 or 99.125 (depends on the method).

(b) IQR = 0.9 or 1 (depending on the method), and no value falls outside ($Q_1 - 1.5$IQR, $Q_3 + 1.5\,IQR$), so there are no outliers. A box plot can be easily constructed from {min, $Q_1$, $\tilde{x}$, $Q_3$, max}.

**Question 6.** (b) For $\alpha = 0.3, 0.4, 0.5$, the prediction for Dec 1997 and the MAPE are, respectively:

{123.57, 2.63%}, {123.37, 2.51%}, {123.10, 2.42%}.
The predictions are calculated recursively using the formula $\mathrm{EWMA}_t = \alpha\, x_t + (1-\alpha)\,\mathrm{EWMA}_{t-1}$. Among these values of $\alpha$, 0.5 is the best as it gives the least MAPE.

**Question 7.** (a) Proportion $q$ of the people will answer question A, and only those who chew gum will answer Yes; proportion $(1-q)$ will answer question $B$, and only those born withing 2 weeks of Christmas will answer Yes.
Assuming that birthdays are uniformly distributed, then the proportion of people born *within 2 weeks* of Christmas is about 1/13. Thus we expect

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

which gives the estimate

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

(b) If $p < (1-q)/13$, then the above estimate becomes negative; if $p > (1+12q)/13$, then the estimate becomes greater than 1.

(c) $q = 0$ 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 $x$, since the expression for $x$ has $q$ in the denominator.

**Question 8.** (a) The population variance is given by

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

(b) There are $4^2 = 16$ random samples in total: $\{1,1\}$, $\{1,2\}$, $\dots$, $\{4,4\}$.
For $\{1,1\}$, $s^2 = 0$; for $\{1,2\}$, $s^2 = 0.5$, etc. After computing $s^2$ for all 16 samples, we find that 0 occurs with frequency 4/16, 0.5 with frequency 6/16, 2 with frequency 4/16 and 4.5 with frequency 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 $\mathrm{E}(s^2) = \sigma^2$.

**Question 9.** $\mathrm{E}(\bar{X}) = (\mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)\big)/n$, which by symmetry equals $\mathrm{E}(X_1)$; here $\mathrm{E}_i$ denotes the number on the $i$th observed tank. Then

$$\mathrm{E}(\bar{X}) = \mathrm{E}(X_1) = \frac{1}{N+1}\big(0+1+2+\cdots+N\big) = \frac{N(N+1)/2}{N+1} = \frac{N}{2}.$$

Note that the *total number* of tanks is actually $N+1$; this is the quantity we want to estimate. By linearity of expectation, $\mathrm{E}(2\bar{X}+1) = 2\frac{N}{2} + 1 = N+1$. Thus $2\bar{X}+1$ is an unbiased estimator for the total number of tanks.