Please use this proforma at the beginning of your TMA to indicate how/if you have used generative AI.

I have used Generative AI in this TMA (such as Copilot, Gemini or ChatGPT) to help with the following:
[please tick all that apply]

☒ As a starting point or inspiration with a part of the TMA.

☐ To improve my own work, like the interpretation/summary of results.

☐ To summarise materials I found on the web for this TMA

☐ I did not use generative AI to help me with this TMA

Q 1.

(a)

(i) If $W = aX + bY$ and $E(X) = E(Y) = \mu$, then

$$\begin{aligned} E(W) &= aE(X) + bE(Y) \\ &= a\mu + b\mu \\ &= (a+b)\mu \end{aligned}$$

(ii) For W to be an unbiased estimator of μ, we must have $(a+b)\mu = \mu$ and therefore the constraint $a + b = 1$ must be satisfied. Since $b = 1 - a$, we can rewrite W as a formula involving a, X and Y only:

$$W = aX + (1-a)Y$$

(b)

(i)

For a pmf to be valid, each individual probability must be positive. By definition this requires $0 < \theta$, and

$$\begin{aligned} \frac{1}{4}\big(1 - 2\theta\big) &> 0 \\ 1 - 2\theta &> 0 \\ 1 &> 2\theta \\ \frac{1}{2} &> \theta \end{aligned}$$

Hence $0 < \theta < \frac{1}{2}$.

(ii)

Let X be a random variable representing the value of the die, and $x_1, x_2, ...x_{1000}$ be observations on the 1000 die rolls. The likelihood of observing this exact set of results is the product of the probabilities of each observation, given the parameter θ:

$$L(\theta) = p(x_1;\theta) \times p(x_2;\theta) \times ... \times p(x_{1000};\theta)$$

Given we know the number of times each value was observed, this can be simplified to

$$L(\theta) = \prod_{i=1}^{6} P(X = i;\theta)^{n_i}$$

where n_i is the number of observed rolls for the ith value of the die. Substituting the pmf and the observed counts gives

$$\begin{aligned} L(\theta) &= \theta^{205} \times \left(\tfrac{1}{4}(1-2\theta)\right)^{154} \times \left(\tfrac{1}{4}(1-2\theta)\right)^{141} \times \left(\tfrac{1}{4}(1-2\theta)\right)^{165} \times \left(\tfrac{1}{4}(1-2\theta)\right)^{145} \times \theta^{190} \\ &= \theta^{395} \times \left(\tfrac{1}{4}(1-2\theta)\right)^{605} \\ &= \left(\tfrac{1}{4}\right)^{605}\theta^{395}(1-2\theta)^{605} \end{aligned}$$

As $\left(\frac{1}{4}\right)^{605}$ is a constant, it can be replaced with C, giving $C\theta^{395}(1-2\theta)^{605}$, as required.

(iii)

Let $L(\theta) = g(\theta)h(\theta)$ with $g(\theta) = C\theta^{395}$ and $h(\theta) = (1-2\theta)^{605}$. Then we have

$$h'(\theta) = 395C\theta^{394},$$

and by the chain rule

$$g'(\theta) = -1210(1-2\theta)^{604}.$$

The by the product rule we have

$$L'(\theta) = 395C\theta^{394}(1-2\theta)^{605} - 1210(1-2\theta)^{604}C\theta^{395}$$

which can be factorised and simplified to

$$\begin{aligned} L'(\theta) &= C\theta^{394}(1-2\theta)^{604}\big(395(1-2\theta) - 1210\theta\big) \\ &= C\theta^{394}(1-2\theta)^{604}(395 - 790\theta - 1210\theta) \\ &= C\theta^{394}(1-2\theta)^{604}(395 - 2000\theta) \end{aligned}$$

as required.

(iv) The maximum likelihood estimate of θ is the value of θ satisfying $L'(\theta) = 0$. The equation

$$0 = C\theta^{394}(1 - 2\theta)^{604}(395 - 2000\theta)$$

is satisfied when $\theta = 0$ and $\theta = \frac{1}{2}$, but as $0 < \theta < \frac{1}{2}$, we must have

$$0 = 395 - 2000\theta$$
$$\theta = 0.1975$$

So based on these data, $\hat{\theta} = 0.1975$. A fair die would have equal probability of observing each face $\theta = \frac{1}{6} \approx 0.1667$, so the unfair die has a higher probability of observing a 1 or a 6 than a fair die.

Q 2.

(a)

(i)

In TMA03 it was established that the PM2_5 variable is not normally distributed. Therefore, despite not knowing the population variance, we calculate an approximate z-interval for mu, which is reasonable given the large sample size of $n = 2696$.

The 90% confidence interval for μ is

$$(\mu^-, \mu^+) = \left(\bar{x} - z\frac{s}{\sqrt{n}}, \bar{x} + z\frac{s}{\sqrt{n}}\right),$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and z is the $q_{0.9}$ quantile of the standard normal distribution. Table 2 in the M248 handbook gives $q_{0.9} = 1.645$. Substituting these values gives

$$(\mu^-, \mu^+) = \left(3.762 - 1.645\frac{3.711}{\sqrt{2696}}, 3.762 + 1.645\frac{3.711}{\sqrt{2696}}\right)$$

$$= (3.644..., \ 3.879...)$$

So the approximate 90% confidence interval for μ is (3.644, 3.880) (to 3 d.p.).

(ii)

If the experiment was repeated many times with the same n and a 90% z-interval calculated for μ each time, we would expect 90% of the constructed intervals to contain μ and therefore we would expect 10% of the intervals to not contain μ. It is not that surprising therefore that μ does not lie inside the calculated interval for this experiment.

(iii)

We would expect $0.95 \times 40 = 38$ of the confidence intervals to contain the true population mean. If an experiment is repeated many times for the same n and a 95% confidence interval of a parameter is calculated for each, we expect 95% of those intervals to contain the true population parameter.

(b)

(i)

The sample standard deviations for the coin1 and coin4 variables are 0.543 and 0.363, respectively. As these differ by less than a factor of three, they satisfy the rule of thumb given in Unit 8.

(ii)

Calculating the 90%, two-sample t-interval (assuming equal variance) in Minitab gives (0.709, 1.551). As the interval is positive and does not contain 0, this suggests later coins contained less silver than early coins.

(iii)

The distribution used is the $t(n_1 + n_2 - 2) = t(14)$ distribution, where n_1 and n_2 are the sample sizes for the coin1 and coin4 variables, respectively.

(iv)

The confidence interval would have been (-1.551, -0.709) (i.e. the same values with their order and sign switched). The order is switched because $\theta = h(\mu) = -\mu$ is decreasing (so $\theta^-, \theta^+) = (h(\mu^+), h(\mu^-)))$, and their sign is switched because $h'(\mu) = 0$ when μ is negative.

Q 3. Below is a poster for a non-technical audience showing how two random samples from a population of numbers can vary.

How two random samples from a population of numbers can vary

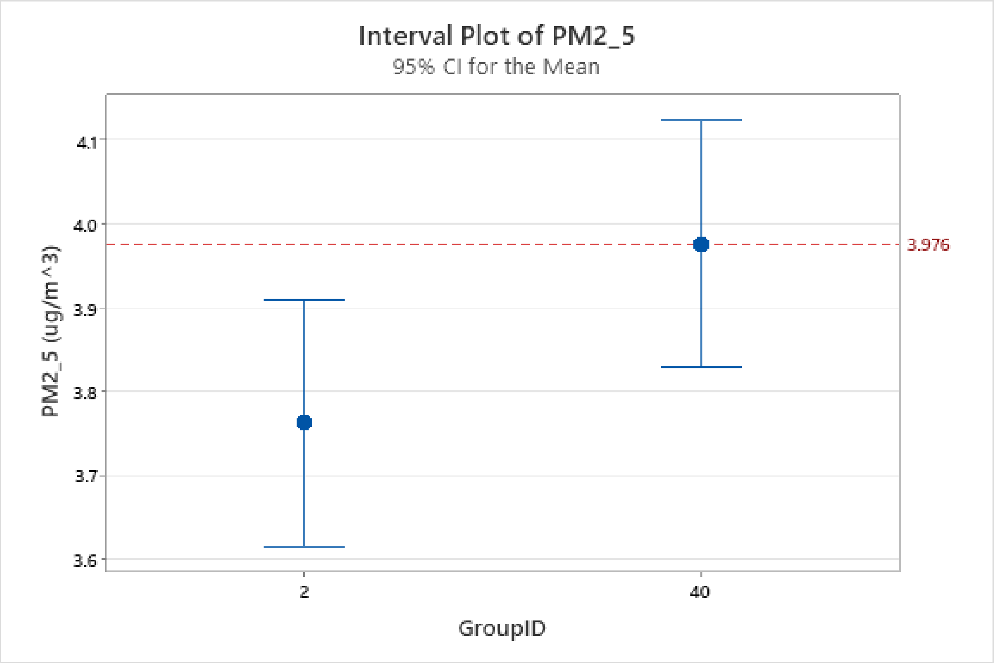

Figure: Sample means (blue dots) and 95% z-intervals of the population mean (crossbars) for each group, and the population mean (red dashed line). Confidence intervals were calculated using pooled variance.

- Open University students were recruited to collect data on atmospheric particulate matter. In total, 107 840 air quality readings were taken and randomly allocated to 40 equally-sized groups, with a GroupID from 1 to 40.

- Of these readings, the PM2_5 variable is of particular interest, being the concentration (in µgm^{-3}) of particles smaller than 2.5 micrometers in diameter in a particular sample.

- The figure above shows the sample means for GroupIDs 2 and 40. Despite being samples from the same population, GroupID=2 has a lower mean (3.762 µgm^{-3}) than GroupID=40 (3.976µgm^{-3}).

- GroupIDs 1 and 40 have similar standard deviations of 3.711 µgm^{-3} and 4.077 µgm^{-3}, respectively, and so pooled variance was used to calculate the z-intervals shown in the Figure.

- Despite being drawn from the same population, only one of the samples' confidence intervals contains the population mean of 3.976 µgm^{-3}. If this was repeated for all samples, we would expect 95% of the intervals to include the population mean.

Source: Open University M248 module website
Produced for TMA04, 15th February 2025