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)

(1)

Let X be the number of Red 1 calls arriving at the LAS in a 30-minute period during daylight hours. As the rate is 3 calls per hour, X can be modeled as $X \sim \text{Poisson}\left(\frac{3}{2}\right)$.

(2)

If $X \sim \text{Poisson}(\lambda)$, then its probability mass function is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Substituting $\lambda = \frac{3}{2}$ and $x = 3$ gives

$$\begin{aligned} P(X = x) &= \frac{e^{-\frac{3}{2}}\left(\frac{3}{2}\right)^3}{3!} \\ &= \frac{0.7530...}{6} \\ &= 0.1255... \end{aligned}$$

So the probability that three Red 1 calls arrive at the LAS in 30 minutes during daylight hours is 0.126 (3 s.f.).

(ii)

(1)

Let Y be the the waiting time (in hours) between the arrivals of two successive Red 1 calls at the LAS during daylight hours. As the rate is 3 calls per hour, Y can be modeled as $Y \sim M(3)$ (an exponential distribution with parameter $\lambda = 3$).

(2)

If $Y \sim M(\lambda)$, then its cumulative distribution function is

$$P(Y \leq y) = F(y) = 1 - e^{-\lambda y}$$

Substituting $\lambda = 3$ and $y = \frac{1}{3}$ gives

$$\begin{aligned} P\left(Y > \tfrac{1}{3}\right) &= 1 - P\left(Y \leq \tfrac{1}{3}\right) \\ &= 1 - F\left(\tfrac{1}{3}\right) \\ &= 1 - \left(1 - e^{-(3\times 1/3)}\right) \\ &= 0.3678... \end{aligned}$$

So the probability that the gap between the arrivals of two successive Red 1 calls at the LAS during daylight hours will exceed 20 minutes is 0.368 (3 s.f.).

(b)

(i)

Minitab calculates the mean and standard deviation of the Interval variable to be 1129 days and 1345 days, respectively. The mean μ and standard deviation σ of an exponential distribution are both $\frac{1}{\lambda}$, where λ is the rate parameter of the distribution. As σ is $\approx 20\%$ larger than μ, this is inconsistent with these being observations from an exponential distribution.

(ii)

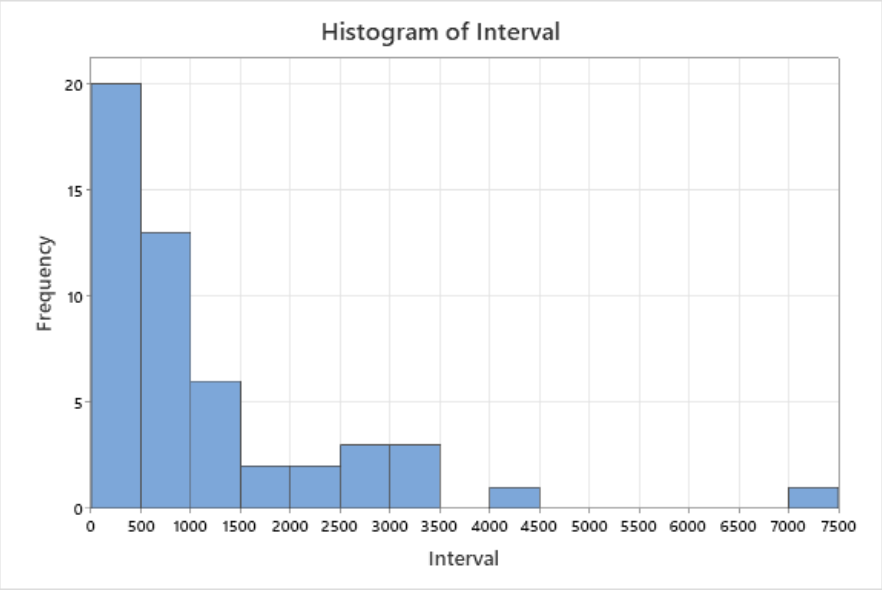

Figure 1: Histogram of the Interval variable.

The shape of the histogram is heavily right-skewed, with a single mode in the first bin. This is consistent with an exponential distribution probability density function, which is a decreasing function.

(iii)

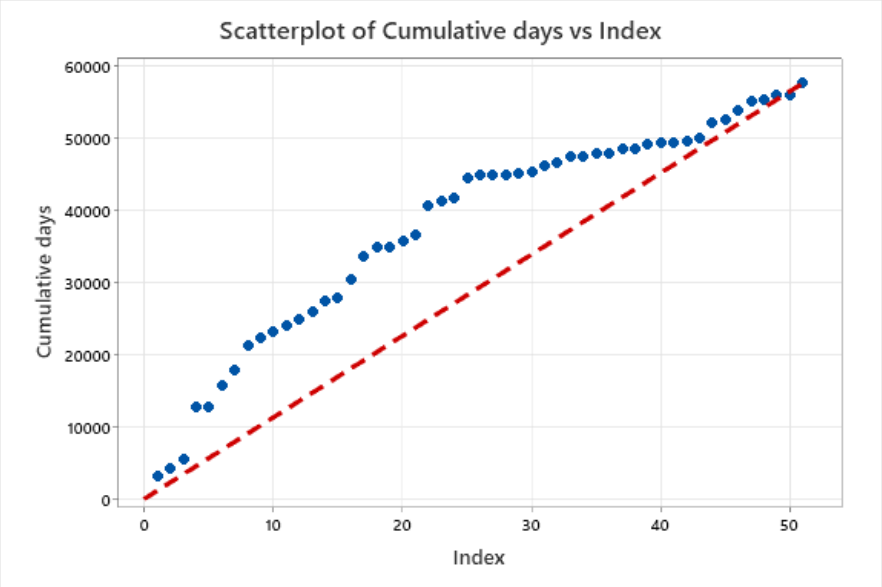

Figure 2: Graph of earthquake index against cumulative sum of the Interval variable. Red line indicates the expected relationship if the time between each earthquake was 1129 days.

Figure 2 shows a non-linear relationship between earthquake index, and the cumulative sum of the Interval variable. If the relationship was approximately linear, this would suggest a constant rate over the period observed (as indicated by the red line). However, the data suggest the rate was highest for earthquakes 1 to 8, slowed from earthquake 9, and slowed again from around earthquake 26, before increasing again from around earthquake 43.

(c)

(i)

The upper quartile is the value $q_{0.75}$ such that

$$F(q_{0.75}) = 0.75 = 1 - (1 - q_{0.75})^2$$

Hence

$$\begin{aligned} 0.75 &= 1 - (1 - q_{0.75})^2 \\ 0.25 &= (1 - q_{0.75})^2 \\ 0.5 &= 1 - q_{0.75} \\ q_{0.75} &= 0.5 \end{aligned}$$

Therefore the upper quartile for this distribution $q_{0.75}$, is 0.5.

(ii)

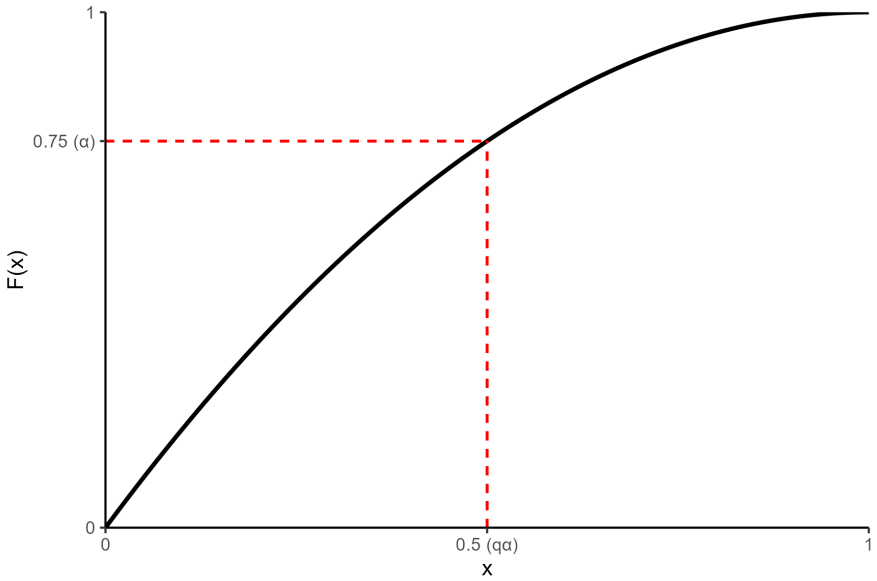

Figure 3: Graph of $F(X)$ with $F(X) = \alpha$ and $X = q_\alpha$ indicated for $\alpha = 0.75$.

Q 2.

(a)

(i)

Given that $X \sim N(98.2, 0.5184)$,

$$P(X \geq 99) = P\left(Z \geq \frac{99 - 98.2}{\sqrt{0.5184}}\right) = P(Z \geq 1.111...),$$

where Z is a standard normal variable. Using the table of probabilities for the standard normal distribution in Table 1 of the M248 handbook (which is accurate to 2 decimal places) gives:

$$\begin{aligned} P(Z \geq 1.11) &= 1 - P(Z < 1.111...) \\ &= 1 - 0.8665 \\ &= 0.1335 \end{aligned}$$

So the proportion of people have a normal body temperature of 99°F or more is 0.134 (to 3 s.f. given the accuracy of Table 1).

(ii)

Let $q_{0.1}$ be the 0.1 quantile of the standard normal distribution. Due to the symmetry around zero of the standard normal distribution, $q_{0.1} = -q_{0.9}$. Table 2 of the M248 handbook gives the value of $q_{0.9}$ for the standard normal distribution as 1.282, so we have $q_{0.1} = -1.282$. Reversing the process of standardisation gives the 0.1 quantile of X as

$$(-1.282 \times \sqrt{0.5184}) + 98.2 = 97.276...$$

So the normal body temperature such that, according to the model, only 10% of people have a lower normal body temperature, is 97.28°F (2 d.p.).

(iii)

For normal random variables W and X, and constants a and b,

$$W = aX + b \sim N(a\mu + b, a^2\sigma^2).$$

where μ and σ are the population mean and standard deviation of X, respectively. Substituting $a = \frac{5}{9}$, $b = \frac{5}{9} \times (-32)$, $\mu = 98.2$, and $\sigma^2 = 0.5184$ gives

$$\begin{aligned} W &\sim N\left(\tfrac{5}{9} \times 98.2 - \tfrac{5}{9} \times (-32), \tfrac{25}{81} \times 0.5184\right) \\ &\sim N(36.777..., 0.16) \end{aligned}$$

So the distribution is $W \sim N(36.78, 0.16)$ (parameters to 2 d.p.).

(b)

(i)

A normal probability plot (Figure 4) is a suitable graph to investigate whether a normal distribution might be a good model. As the points do not fall along a diagonal line, this suggests a normal probability distribution is *not* a plausible model for these data.

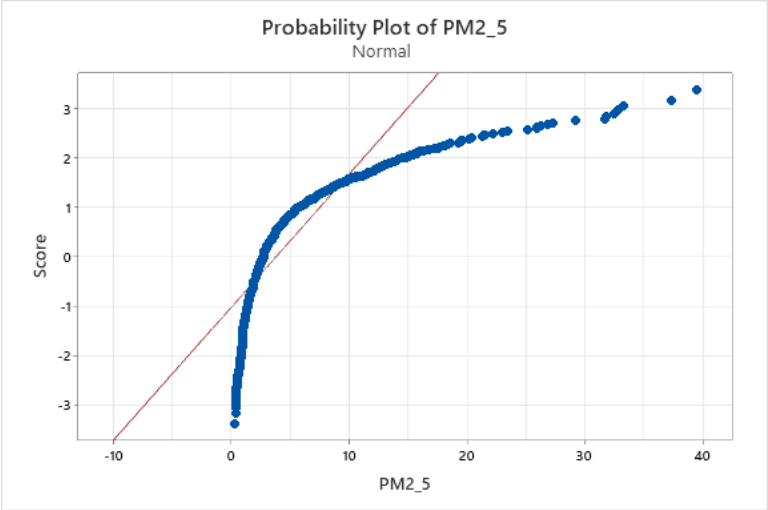

Figure 4: Normal probability plot of PM2_5 for recordings with GroupID = 2.

(ii)

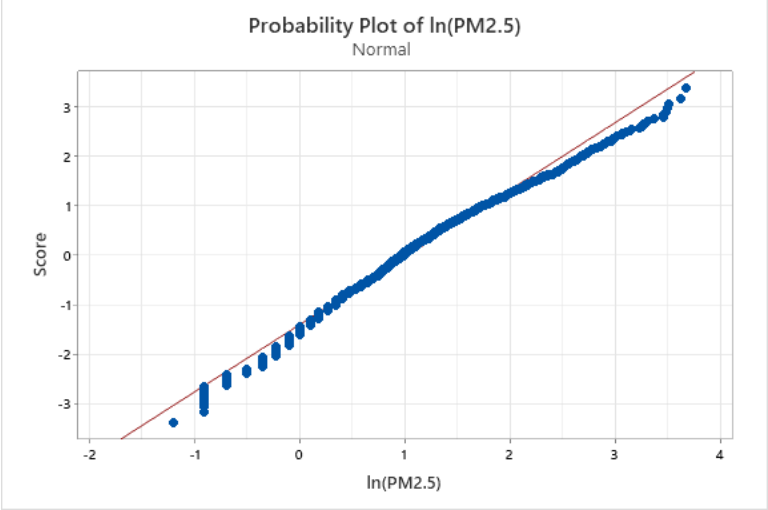

Figure 5: Normal probability plot of ln(PM2_5) for recordings with GroupID = 2.

As the points *do* fall along a diagonal line, this suggests a normal probability distribution *is* a plausible model for ln(PM2_5). As the population of air quality data was randomised into different groups, I would expect the ln(PM2_5) variable in GroupID=40 to have a similar distribution to GroupID=2.

(c)

(i)

The central limit theorem tells us that the sample mean $\bar{X}$ has the approximate distribution

$$\begin{aligned}\bar{X} &\approx N\left(\mu, \frac{\sigma^2}{n}\right) \\ &\approx N\left(3.976, \frac{18.460}{2696}\right) \\ &\approx N(3.976, 0.007)\end{aligned}$$

(both to 3 d.p.).

(ii)

Let $\bar{X}$ represent the sample mean as a random variable. Using the parameters of the sampling distribution identified in the previous answer, we begin by standardising the value of 3.762:

$$\begin{aligned}P(\bar{X} < 3.762) &= P\left(Z < \frac{3.762 - 3.976}{\sqrt{0.0068...}}\right) \\ &= P(Z < -2.59)\end{aligned}$$

where Z is a standard normal variable. Using Table 1 of M248 Handbook, we find the probability as

$$\begin{aligned}P(Z < -2.59) &= 1 - P(Z < 2.59) \\ &= 1 - 0.9952 \\ &= 0.0048\end{aligned}$$

(to 2 d.p.).

We perform the same procedure as above, for the quantile 3.976:

$$\begin{aligned}P(\bar{X} < 3.976) &= P\left(Z < \frac{3.976 - 3.976}{\sqrt{0.0068...}}\right) \\ &= P(Z < 0) \\ &= 0.5\end{aligned}$$

(iii)

Running commentary TMA01 Q1b: Description, Similarities and Differences for Group 2 and Group 40.

As shown in Figure 1 (TMA01), the PM2.5 variable has a similar central location in Groups 2 and 40, with the mean and median slightly higher in Group 40. Group 40 has a slightly higher standard deviation and, as shown in Figure 2 (TMA01), interquartile range. Both groups have the same minimum value of 0.3, but Group 40's maximum value is more than twice the largest value in Group 2. Figure 2 (TMA01) suggests both data distributions exhibit positive skew, with Group 40's skew being slightly more pronounced.

The number of low PM2.5 values in a sample of 10 readings from Group 2 is fewer than expected and in a sample of 10 readings from Group 40 is fewer than expected.

The PM2_5 variable is not normally distributed, whereas the ln(PM2_5) variable *is* normally distributed. Both of these observations hold for Group 2 and Group 40. As the probability of observing a sample mean smaller than that of Group 2 is extremely low, this suggests the mean of PM2_5 for Group 2 is near the lower end of the distribution of means. However, the sample mean for Group 40 is equal to the expected population mean, which is why the probability of observing a sample mean lower than this is exactly half.