Statistics [Q.1] [20 marks]

a) **[6 Marks]**

i) X = number of yellow cars in 15 min period $\sim$ Poisson$(4.8/4)$ = **Poisson(1.2)** [**1 mark**]
$P(X \geq 1) = 1 - P(X = 0) = 1 - 1.2^0 \exp -1.2/0! = \mathbf{0.6988.}$ [**1 mark**]

ii) Y = number of times "S or S!" declared in 10 breaks $\sim$ **Binomial(10,0.6988)** [**1 mark**]
$P(Y \geq 9) = P(Y = 9) + P(Y = 10) = 0.1197 + 0.0278 = \mathbf{0.1474}$ [**1 mark**]

iii) $Z = 5Y$ [**1 mark**]
$E(Z) = 5E(Y) = 5n\pi = 5 * 10 * 0.6988 =$ **34.94 minutes** [**1 mark**]
$Var(Z) = 25Var(Y) = 25n\pi(1 - \pi) = 25 * 10 * 0.6988 * (1 - 0.6988) = 52.61$ minutes
$sd(Z) = \sqrt{52.51} =$ **7.25 minutes** [**1 mark**]

b) **[7 marks]**

i) **[3 marks]**
$\hat{\pi} = 0.27$ [**1 mark**], $z_{0.95} = 1.645$ [**1 mark**] and so
standard error $= z_{0.95}\sqrt{\hat{\pi}(1 - \hat{\pi})/n} = 0.07303$
giving a 90% CI for π of $\hat{\pi} \pm z_{0.95}\sqrt{\hat{\pi}(1 - \hat{\pi})/n} =$ **(0.1970,0.3430)** [**1 mark**].

ii) **[3 marks]**
[1 mark per correct assumption, 1/2 mark for reasonable attempt to check]
– CLT empirical rule $n\hat{\pi}(1 - \hat{\pi}) > 5$ (here 100*0.27*(1-0.27)=19.71>5 is ok)
– independent observations with the same probability π (may not be reasonable as the tweets may be on related or different subjects, and be close or distant in time)

c) **[7 marks]**

i) Hypothesis: $H_0 : \mu = 100 (= \mu_0)$ versus $H_1 : \mu > 100$ [**1 mark**]

ii) If H_0 is true then $t_0 = (\bar{x} - \mu_0)/(s/\sqrt{n}) \sim t_{n-1} = t_{49}$ [**1 mark**]

iii) $t_0 = (105.3 - 100)/(15.3/\sqrt{50}) = 2.45$ [**1 mark**]

iv) $p = P(t_{n-1} > t_0) = P(t_{49} > 2.45)$ [**1 mark**]

v) Using t_{50} on tables we have $0.005 < p < 0.01$ [**1 mark**]

vi) $p < 0.01$ therefore we reject H_0 in favour of H_1 [**1 mark**]

vii) i.e. The true mean battery duration is greater than 100 hours. [**1 mark**]

Note: The above marks for a "p-value" solution. If doing a "test statistic/rejection region" solution then replace iv)–vi) with:

iv) For a 1% significance level reject if $t_0 > F_{crit}$ **where** $F_{crit} = F_{n-1,0.99}$ [**1 mark**]

v) From tables $F_{49,0.99} \approx F_{50,0.99} =$ **2.403** [**1 mark**]

vi) $t_0 > F_{crit}$ therefore we reject H_0 in favour of H_1 [**1 mark**]

Q2. [**20 marks**]

(a) *[3 marks, 1 mark for each point.]*

- *Comment about location:* The scores are highest for temperatures $120°C$ and $140°C$; in the middle for $100°C$; and lowest for $160°C$.
- *Comment about spread:* The observed variability is quite similar for all four temperatures.
- *Comment about shape/outliers:* The distribution of scores for each temperature is fairly symmetrical and no outliers are present.

[Note: the sample size is quite small here.]

(b) *[3 marks, 1/2 mark for each assumption and 1/2 mark for an appropriate comment.]*

- The observations for the scores for each temperature were drawn from Normal distributions; there is no way of checking this here (a quantile/qq-plot would be needed, a symmetrical boxplot does not tell anything about normality).
- The observations are independent; there is no way of checking this here.
- The variances of the scores of each temperature are the same; using the rule-of-thumb (i.e., the ratio of the largest sample standard deviation to the smallest one is smaller than 2), this assumption is not-acceptable here, although the sample size is very small.

(c) *[3 marks, 1/2 mark for each missing value in the table.]*

$$\begin{aligned}
(1) &= k - 1 = 3\\
(2) &= 484.25 - (5) = 434.25\\
(3) &= (2)/(1) = 144.75\\
(4) &= (6) - (1) = 8\\
(5) &= (4) \times 6.25 = 50.00\\
(6) &= n - 1 = 11
\end{aligned}$$

This yields the full table:

Source	df	SS	MS	F
Treatment	$(1) = 3$	$(2) = 434.25$	$(3) = 144.75$	23.16
Error	$(4) = 8$	$(5) = 50.00$	6.25	
Total	$(6) = 11$	484.25		

[Note: please pay attention to carry-over mistake and rounding errors. For example, if (1) is wrong but (3) is calculated as 434.25/(1), the 1/2 mark for (3) should be granted.]

(d) *[5 marks, 1 mark for each of the following points.]*

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. H_a : not all the means are equal (an alternative hypothesis stated as $H_a : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ is obviously not correct).
- Rejection criterion: reject H_0 if $f_0 > f_{3,8;0.95} = 4.066$ (MATLAB), from the stats table, $f_{3,8;0.95} = 4.07$.
- The observed value of the test statistic is $f_0 = 23.16$, which is larger than $4.066 \rightarrow$ reject H_0.
- The p-value is $p = P(F_{3,8} > 23.16)$. From the table, it can be concluded that $p < 0.005$.
- Conclusion: there is very strong evidence that the scores for each temperature are not all the same.

(e) *[4 marks; 1 mark for a correct expression of the CI, 2 marks for the correct values and 1 mark for a correct conclusion on H_0 : $\mu_1 = \mu_2$.]*

A 95% confidence interval for $\mu_1 - \mu_2$ is:

$$\left[(\bar{x}_1 - \bar{x}_2) \pm t_{n-k;1-\alpha/2}\sqrt{\mathrm{ms}_{\mathrm{Er}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right]$$

$$= \left[(44.33 - 54.33) \pm 2.306\sqrt{6.25 \times \left(\frac{1}{3} + \frac{1}{3}\right)}\right]$$

$$= [-14.70, -5.29].$$

[Note: in the ANOVA context, students were asked to always use ms_{Er} as estimate of the common standard deviation σ, that's why $t_{n-k;1-\alpha/2} = t_{8;0.975} = 2.306$ is used as critical value.

Consequently, answers using a classical two-sample t-confidence interval (hence based on a $t_{n_1+n_2-2}$ sampling distribution) are not totally correct. Half marks (1/2) may still be granted if all but this is correct.]

Conclusion: 0 does <u>not</u> belong to that 95% confidence interval for $\mu_1 - \mu_2$, which indicates that there is a significant difference (at 5% significance level) between μ_1 and μ_2.

(f) *[2 marks, 1 mark for mentioning (or effectively using) a Bonferonni adjustment and 1 mark for a correct conclusion.]*

To relate this to an overall test for $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ at significance level $\alpha = 0.05$, we should compare that p-values to $\alpha/6 = 0.00833$ (Bonferonni adjustment).

Some p-values from the table are obviously smaller than $\alpha/6 = 0.00833$ and so we would reject the null hypothesis that they are the same. Others (such as $100°C$ vs $140°C$, $100°C$ vs $160°C$ and $120°C$ vs $140°C$) have p-values that > 0.00833, so we would not reject the null hypothesis and so the conclusion will be different from (d).

Statistics

Q3. [**20 marks**]

(a) i. [1 mark] `Number of Foals` $= -1.91793 + 0.15862($`Number of Adults`$)$

ii. [1 mark] $r^2 = 0.862$

(b) [1 mark] $\hat{\beta}_1 = 0.15862$

(c) [1 mark] $r = \sqrt{r^2} = \sqrt{0.862} = 0.928$

(d) [1 mark] $s = \hat{\sigma} = 4.487$

(e) [6 marks total: 2 marks for hypotheses, 1 mark for df, 2 marks for rejection region (or 1 mark for observed test statistic, 1 mark for p-value), 1 mark for conclusion (deduct 0.5 mark if the conclusion is not related to the original problem)]

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ [2 marks: 1 mark for H_0 and 1 mark for H_a]
- $t_{40,0.975} = 2.021$.
  Rejection criterion: Reject H_0 if

  $$\hat{\beta}_1 \notin \left[-t_{40,0.975}\frac{s}{\sqrt{s_{xx}}}, t_{40,0.975}\frac{s}{\sqrt{s_{xx}}}\right] = [-2.021 \times 0.01004, 2.021 \times 0.01004] = [-0.0203, 0.0203] .$$

  Or, $t = \dfrac{0.15862}{0.01004} = 15.807$, $df = 40$. $p - value = 2P(T > 15.807) = 8.44e - 19$. (or $p - value = 2P(T > 15.807) < 0.001$ from the table)
- Reject H_0. `Number of Adults` is significantly associated with `Number of Foals` (or something similar that ties original problem with statistical results)

(f) [2 marks: 1 mark for correct t quantile, 1 mark for correct interval]
$t_{40,0.95} = 1.684$

$$0.15862 \pm 1.684 \times 0.01004 = [0.1417, 0.1755]$$

(Alternatively, $z_{0.95} = 1.645$. By CLT (large n), the approximate CI is $0.15862 \pm 1.645 \times 0.01004 = [0.1421, 0.1751]$. However, to use the approximate CI, **CLT** must be mentioned. If **CLT** is mentioned, award 2 marks. Otherwise, award 0 mark.)

(g) [4 marks: 1 mark for $\hat{y}(x_0)$, 1 mark for t quantile, 1 mark for correct equation, 1 mark for correct interval]

i. $\hat{y}(x_0) = -1.91793 + 0.15862(120) \simeq 17$ [1 mark]

ii. $t_{40,0.995} = 2.704$

$$\begin{aligned}
& \hat{y}(x_0) \pm s t_{n-2,1-\alpha/2}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}} \\
= \; & 17 \pm 4.487 \times 2.704\sqrt{1 + \frac{1}{42} + \frac{(120 - 4738/42)^2}{199970.5}} \\
= \; & [4.722, 29.278]
\end{aligned}$$

(h) i. [2 marks: 1 mark if one assumption is correct, 1.5 marks if two assumptions are correct]

- $e_i's$ have been drawn independently of one another
- $e_i's$ have the same variance
- $e_i's$ have been drawn from a normal distribution

ii. [1 mark: 0.5 mark for comment on residual plot, 0.5 mark for comment on residual plot] Residual plot does not show obvious pattern. QQ plot is closed to a straight line.