[redacted]

(1) The University [redacted] wants to evaluate how many employees have returned to campus after the easing of Covid restrictions. A survey of 45 randomly selected lecturers within the School of Mathematics found that 32 had returned to campus whilst the others continue to work from home.

State whether each of the following statements is either ‘true’, ‘false’ or ‘may be true or false’. Fully justify your answer.

(a) In a statement, the University Chancellor says that 29% of all university employees are working from home. **[4 marks]**

(b) Let $\hat{p}$ be the sample proportion of mathematics lecturers who have returned to campus. In statistics, the uncertainty in this estimate comes from the fact that the true proportion, p, is unknown. **[4 marks]**

(c) The calculation $1 - \hat{p}^6$ represents the maximum likelihood estimate for the probability that at least one employee on a corridor of 6 mathematics lecturers is working from home. **[4 marks]**

(d) By assuming that the survey sample size is sufficiently large, an approximate confidence interval for p is (63%, 88%) rounded to the nearest whole percent. **[4 marks]**

(e) The head of the school claims that that the actual proportion of returned lecturers in the school is 80% or higher. To support this claim, the head of the school has provided the following calculation and states that this p-value is smaller than the 5% significance level.

```
dbinom(32, size = 45, prob = 0.8)
[1] 0.04738374
```

[4 marks]

[Please turn over]

(2) Let $X_1, \ldots, X_n$ denote independent and identically distributed normal random variables with expectation μ and variance σ^2.

(a) Assume n is even and consider the estimator $\tilde{X}$ defined below where the even indexed samples are given twice the weight of the odd indexed samples:

$$\tilde{X} = \frac{2}{3n} \sum_{i=1}^{n/2} (X_{2i-1} + 2X_{2i}).$$

(i) Show that $\tilde{X}$ is unbiased in estimating μ. **[4 marks]**

(ii) Prove that $\mathrm{Var}(\tilde{X}) = \frac{10}{9n}\sigma^2$ and that $\tilde{X}$ is a consistent estimator for μ. **[4 marks]**

(iii) Derive a $100(1-\alpha)\%$ confidence interval for μ based on the estimator $\tilde{X}$, assuming that σ^2 is known. Clearly state any distributional results that you are using and carefully define any notation used. **[4 marks]**

(b) A general estimator for the population variance is given by:

$$S_k^2 = \frac{1}{k} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

for some constant k, where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

(i) Derive an expression for the expectation for S_k^2. **[2 marks]**

(ii) State the value for k so that:

(A) S_k^2 is unbiased in estimating σ^2? **[1 mark]**

(B) S_k^2 is the maximum likelihood estimate for σ^2? **[1 mark]**

(iii) Is S_k^2 a consistent estimator for σ^2 when $k = \frac{n^2-1}{n^2}$? **[2 marks]**

(iv) Let $S_k = \sqrt{S_k^2}$ be a general estimator for the standard deviation σ. Is S_{n-1} unbiased? Justify your answer. **[2 marks]**

[Please turn over]

(3) A behavioural scientist is researching how people communicate with each other. She designs an experiment where two people must work together to create a statue out of building blocks from an image, but the image can only be seen by person 1 and must give instructions to person 2 who can only build. The experiment consists of two types of pairing; pairing A consists of two siblings and pairing B where the couple are strangers. The scientist scores each couple on how much of their communication is constructive towards the task. Assume that the scores are normally distributed. Below is a summary of the experimental data:

$$\begin{array}{llll} \text{Pairing A (siblings):} & n_A = 10 & \bar{x}_A = 76.6 & s_A^2 = 21.6 \\ \text{Pairing B (strangers):} & n_B = 10 & \bar{x}_B = 67.5 & s_B^2 = 112.5 \end{array}$$

Note: The R extract at the end of the question contains some useful quantile values.

(a) Perform a two-sample t-test at the 5% level to investigate whether mean communication scores, μ_A and μ_B, are equivalent. Carefully follow the 5 steps of performing a hypothesis test. **[8 marks]**

(b) An assistant runs the following R code where the vectors `xA` and `xB` contain the data from pairings A and B respectively. What type of test has the assistant performed? Explain why it is not appropriate for this experiment. **[2 marks]**

```
> t.test(xA - xB, alternative = "two.sided", conf.level = 0.95)
data:  xA - xB
t = 1.898, df = 9, p-value = 0.09018
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.746042 19.946042
```

(c) Existing research suggests that communication between family members is more effective and this would naturally result in expected communication scores 5 units higher than between strangers. Re-assess the two-sample t-test performed in part (a) to investigate the hypotheses at the 5% significance level:

$$H_0 : \mu_A - \mu_B = 5 \quad \text{vs} \quad H_1 : \mu_A - \mu_B > 5$$

Is there evidence to suggest that the experiment helps in forming better constructive sibling communication compared to that between strangers over and above the known family effect? **[4 marks]**

(d) Why it is advisable to perform an F-test before conducting a two-sample t-test? **[2 marks]**

(e) Perform an F-test at the 5% significance level. Provide a comment about the reliability of the conclusions made to the above test. **[4 marks]**

```
> ## Student's t-quantiles       ## F-distribution quantiles
> qt(c(0.95, 0.975), df=20)      > qf(c(0.05, 0.95), df1=10, df2=10)
[1] 1.724718 2.085963            [1] 0.3357691 2.9782370
> qt(c(0.95, 0.975), df=19)      > qf(c(0.025, 0.975), df1=10, df2=10)
[1] 1.729133 2.093024            [1] 0.2690492 3.7167919
> qt(c(0.95, 0.975), df=18)      > qf(c(0.05, 0.95), df1=9, df2=9)
[1] 1.734064 2.100922            [1] 0.3145749 3.1788931
> qt(c(0.95, 0.975), df=9)       > qf(c(0.025, 0.975), df1=9, df2=9)
[1] 1.833113 2.262157            [1] 0.2483859 4.0259942
```

[Please turn over]

(4) Ten drivers are taking part in a road safety investigation where each drives a vehicle around a track at any speed and then, at five random occurrences, must perform an emergency stop when instructed to by the investigator. The speed of the vehicle at the initiation of the emergency stop is measured accurately with a digital speedometer, but the distance the vehicle then travels until coming to a complete stop is measured with error. The investigators wish predict the stopping distance given the vehicle's speed and has provided the following summary statistics:

	Mean	Std. Dev.	Min.	Max.
Vehicle Speed (km/h)	48.58	6.82	35	65
Stopping Distance (meters)	81.76	20.61	39	132
	$n = 50$, Correlation = 0.914			

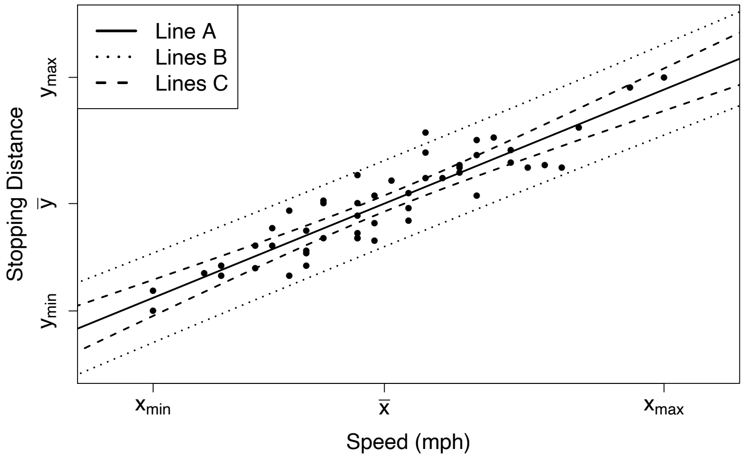
(a) Compute the least squares estimates and write the equation for the fitted regression line. Clearly explain any notation that you are using. **[4 marks]**

(b) Based on the description of how the data was collected, which of the linear regression assumptions may be violated? **[2 marks]**

(c) The scatter plot presented below depicts the fitted regression line (Line A) and two other pairs of lines.

(i) Explain what the outer pair of dotted lines (Lines B) represent. **[2 marks]**

(ii) The pair of dashed lines (Lines C) visibly curve away from the fitted regression line. State the formula for these lines and explain what is causing the observed curvature. **[4 marks]**



(d) The road safety investigators also collected data on the depth of the tyre tread. Below is an incomplete summary of the fitted model $\mathbb{E}[Y_i] = \alpha + \beta_1 x_i + \beta_2 t_i$ where t_i denotes the additional tyre tread depth explanatory variable.

```
Coefficients:
            Estimate Std. Error t value
(Intercept) -36.5879     9.0722  -4.033
Speed         2.8774     0.1629  17.664
TyreTread    -5.9344     1.7103  -3.470
---
Residual standard error: 7.613
Multiple R-squared:  0.8691,    Adjusted R-squared:  0.8635
F-statistic:   [?] on [?] and [?] DF,  p-value: < 2.2e-16
```

(i) Use the values presented R extract and the summary statistics table to calculate the three missing values for the F-statistic indicated by “[?]”.

[4 marks]

(ii) Perform an appropriate hypothesis test at the 5% significance level to assess whether stopping distance depends on tyre tread depth. State clearly the null and alternative hypotheses. You may use the following information.

[4 marks]

```
> qnorm(0.95)                  > qnorm(0.975)
[1] 1.644854                   [1] 1.959964
> qt(0.95, df = 48)            > qt(0.975, df = 48)
[1] 1.677224                   [1] 2.010635
> qt(0.95, df = 47)            > qt(0.975, df  = 47)
[1] 1.677927                   [1] 2.011741
```

(5) A textile company is investigating what effect the yarn tension and feed speed has on the durability of knitted clothing. The knitting machine has three tension settings and four speed settings. For each combination of the settings, five fabric samples were made and their durability were assessed by measuring how much pressure the fabric could resist until the material ripped. The table below gives the mean pressure (in kPa) within each of the 12 settings.

	slow ← Speed → fast			
Tension	I	II	III	IV
Loose	702	702	713	649
Normal	691	723	720	647
Tight	707	711	742	702

Below is an extract of an analysis performed in Rstudio where `rip_pressure_data` contains the experimental data.

```
> model <- lm(pressure ~ tension + speed, data = rip_pressure_data)
> anova(model)
Analysis of Variance Table

Response: pressure
          Df Sum Sq Mean Sq F value    Pr(>F)
tension    2   6668  3333.8  2.6241 0.0817310
speed      3  28841  9613.8  7.5674 0.0002582
Residuals 54  68602  1270.4
```

(a) What type of ANOVA investigation has been performed in the above extract? **[2 marks]**

(b) State the assumptions that underpin this ANOVA model. **[4 marks]**

(c) In words, state the null and alternative hypotheses only for the tension settings. **[3 marks]**

(d) Conclude both hypothesis tests at the 5% significance level. Ensure that you explain how you arrive at your decisions. **[2 marks]**

(e) Where there is evidence of difference in the tension/speed settings, compute the least significant difference and identify where the statistical differences exist. **[6 marks]**

The following `R` output may be useful:

```
> qt(0.95, df = 3)        > qt(0.975, df = 3)
[1]  2.353363            [1] 3.182446
> qt(0.95, df = 4)        > qt(0.975, df = 4)
[1]  2.131847            [1] 2.776445
> qt(0.95, df = 54)       > qt(0.975, df = 54)
[1]  1.673565            [1] 2.004879
> qt(0.95, df = 60)       > qt(0.975, df = 60)
[1] 1.670649             [1] 2.000298
```

(f) Without performing any further calculations, explain how your approach to the ANOVA investigation in parts (d) and (e) would change if a 10% significance level was instead specified. **[3 marks]**

[End of Paper]