Section A

QUESTION 1

Answer the following questions with either TRUE or FALSE

(a) The Type I Error Rate is the probability we fail to reject the null hypothesis if it is not true
FALSE (1 mark)

(b) In logisitic regression we assume that $\log(p_i/(1-p_i)) = \beta_0|\beta_1 + \epsilon_i$
FALSE (1 mark)

(c) Increasing the sample size for a hypothesis test increases the power of that test.
TRUE (1 mark)

(d) A 95% confidence interval doesn't provide any information about the true parameter value.
FALSE (1 mark)

Section B

QUESTION 2

The Acme Tyre Company has three plants A, B, and C that produce tyres; they produce respectively 43%, 21% and 36% of the tyres produces by The Acme Tyre Company. Each plant produces a certain number of defective tyres that must be recycled rather than sold. The defect rates for the three plants are respectively: Plant A 0.01, Plant B 0.07, Plant C 0.03

(a) What is the What proportion of all the tyres produced by The Acme Tyre Company are recycled? (2 marks)

$$\begin{aligned} Pr(R) &= Pr(R|A)Pr(A) + Pr(R|B)Pr(B) + Pr(R|C)Pr(C) \\ &= (0.01)(0.43) + (0.07)(0.21) + (0.03)(0.36) \\ &= 0.0298 \end{aligned}$$

(b) What is the probability that a given tyre produced by the Acme Tyre Company sold at Bob James Tyre Mart was produced at Plant B? (2 marks)

$$\begin{aligned} Pr(B|NR) &= \frac{Pr(NR|B)Pr(B)}{Pr(NR)} \\ &= \frac{(0.93)(0.21)}{1 - 0.0298} \\ &= 0.20 \end{aligned}$$

QUESTION 3 Given the probability distribution for $1 < X < \infty$

$$f(x|\theta) = \frac{\theta}{x^{\theta+1}}$$

(a) What is the MLE for θ? (2 marks)

$$\hat{\theta} = \frac{n}{\sum \log(x_i)}$$

(b) The MLE for θ is a function of $T(\boldsymbol{x})$ a sufficient statistic of the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, for the MLE of θ what is $T(\boldsymbol{x})$? (2 marks)

$$T(\boldsymbol{x}) = \sum \log(x_i)$$

QUESTION 4

Given a sample of data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ of size $n = 297$ from $X \sim \text{Binomial}(n, p)$.

(a) What are mean and variance? (2 marks)

$$E(X) = np$$

$$\text{Var}(X) = np(1 - p)$$

(b) If $\bar{X} = 0.73$ what is the 95% confidence interval for p? (2 marks)

$$\begin{aligned} \text{CI} &= \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= (0.68, 0.78). \end{aligned}$$

QUESTION 5

The linear regression problem can be written as:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

What are the sampling distributions for the parameters β_0 and β_1

(3 marks)

They follow t distributions, e.g.

$$\frac{\hat{\beta}_0 - \beta_0}{s_{\beta_0}} \sim t_{n-p-1}$$

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} \sim t_{n-p-1}$$

QUESTION 6

A Hypothesis relies on a set of hypotheses, a test statistic, the sampling distribution of the test statistic, and a rejection region for that test statistic.
Given a scenario where you wanted to test the that the mean of a Poisson random variable was $\lambda = 3$:

(a) What are the hypotheses? (1 mark)
$H_0 \colon \lambda = 3$ **and** $H_A \colon \lambda \neq 3$

(b) What is the test statistic given one trial? (1 mark)
X

(c) What is the sampling distribution of the test statistic (1 mark)
Poisson

(d) If the Type I error rate is set to $\alpha \leq 0.1$ define a rejection region. (2 marks)
$X < 0 \cup X > 6$

Section C

QUESTION 7

Fuel efficiency in auto-mobiles can be influences by a number of characteristics. See the linear regression output below and answer the following questions.
Results of linear regression analysis are shown below:

```
Call:
lm(formula = mpg ˜ ., data = auto_mpg)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6927 -2.3864 -0.0801  2.0291 14.3607

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.454e+01  4.764e+00  -3.051  0.00244 **
cyl         -3.299e-01  3.321e-01  -0.993  0.32122
disp         7.678e-03  7.358e-03   1.044  0.29733
hp          -3.914e-04  1.384e-02  -0.028  0.97745
gvw         -6.795e-03  6.700e-04 -10.141  < 2e-16 ***
accel        8.527e-02  1.020e-01   0.836  0.40383
year         7.534e-01  5.262e-02  14.318  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.435 on 385 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.8093,Adjusted R-squared:  0.8063
F-statistic: 272.2 on 6 and 385 DF,  p-value: < 2.2e-16
```

(a) What proportion of observed variance in fuel efficiency is explained by the covariates? (1 mark)

80.93%

(b) What is the 95% confidence interval for β_5, the coefficient for acceleration? (2 marks)

(-0.115,0.285)

(c) Is there evidence to reject the assertion that displacement has no effect on fuel efficiency? Why or Why Not? (2 marks)

No, p-value for test of $H_0 : \beta_2 = 0$ is 0.29733

(d) There are several variables in the model that are not statistically significant, should these be removed from the analysis?

It depends, we should investigate for multicollinearity or correlation between the covariates, and possible remove variables that are highly correlated.

(e) What are any advantages or disadvantages to keeping variables in the model that are not statistically significant?
Keeping them in the model might improve predictions made from the model, getting rid of them will improve the power of the statistical tests for other variables

(f) What will happen the R^2 value when we add variables that are not statistically significant? What will happen the Adjusted R^2 values?
Typically the $\mathbf{R}^2$ values will increase as we add explanatory variables even if they are not statistically significant. Typically, the same thing will decrease the Adjusted $\mathbf{R}^2$.