Section A

1. Multiple choice: give answers (as a letter A, B, C, D or E only) to each of the 9 questions in your script book. There is no need to show any working.

   (a) C [sqrt(0.25*(1 - 0.25)/200)]
   (b) A (offspring)
   (c) B (= 12.5, 3.31)
   (d) D (res = y - y.hat)
   (e) D
   (f) A
   (g) E
   (h) A
   (i) C

[$2+2+2+2+2+2+2+2+2 = 18$ marks]

Section B

2. (a)
   Number of "successes" in n independent trials,
   Each with probability of success p.
   Example.

   (b)
   Parameters relate to the population.
   Estimates are calculated from a sample (i.e. from data).
   Parameters are fixed, but unknown.
   Examples.

   (c)
   Used to assess whether the data are consistent with a normal distribution,
   Especially residuals.
   Plotted against theoretical quantiles.
   Look for closeness to straight line.
   Example.

   (d)
   An interval within which we're confident that the true value of a parameter lies.
   In the long run, the designated percentage of CIs will contain the true value.
   Larger % $\Rightarrow$ wider interval.
   Example.
   Connection with hypothesis testing.

   (e)
   Take a large number of samples and calculate an estimate.
   Usually applied to the sample mean.
   Example (such as taking samples of size 15 from an exponential distribution).
   Example of plot of distribution.

[$3+3+3 = 9$ marks]

3. **(1)** using 20 CD players once each (all players of the same make and model);
   i. CD player.
   ii. 5 CD players randomly assigned to each of the four brands of battery.
   iii. C: one-way ANOVA

   (2) using 5 CD players four times each (all players of the same make and model);
   i. CD player by battery combination (CD players are blocks)
   ii. One of each brand of battery, assigned to a randomly selected CD player
   iii. D: two-way ANOVA

   (3) using just one CD player, 20 times.
   i. Battery
   ii. Using five of each brand of battery, randomize the order
   iii. C: one-way ANOVA

   (a) 2, with CD player as a block. This allows us to both estimate and absorb player-to-player variation, focusing on the battery differences. 2 points for (3), which eliminates player-to-player variation, but leaves us ignorant of it!

[3 + 3 + 3 + 4 = 13 marks]

Section C

4. (a) i. The χ^2 test statistic.
   ii. 3, 1, 2;
   iii. expected frequency, $e < 5 \ldots$ for the Control.High & Alzheimers.High cells.

   (b) That there is no association between Alzeihmers and Aluminium antacids.

   (c) $\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(213-213.676)^2}{213.676} + \frac{(42-41.324)^2}{41.324} + \frac{(211-210.324)^2}{210.324} + \frac{(40-40.676)^2}{40.676} = 0.0018011.$

   (d) Alz & Al.use (Y/N) are not related; but Alz may be related to the extent of Al.use (given that it is used).

[3 + 1 + 2 + 3 = 9 marks]

5. Algae

   (a) $H_0 : \mu_d = 0; H_1 : \mu_d \neq 0$

   (b) μ_d is the mean of the "Pre weight subtracted from the post weight" (algae treatment) for the population.

   (c) Assume: constant variance, normal distribution, independence, sample is representative, and model is appropriate.

   (d) $s_{\bar{d}} = 1.52$ and $se_{\bar{d}} = 1.52/\sqrt{195} = 0.109$

   (e) $t = 0.37/0.109 = 3.39$

   (f) i) $t(= 3.39) > t_{0.975,194}(= 1.972)$, thus reject the null hypothesis.
   ii) df = 194

   (g) The mean of the "Pre-treatment weight subtracted from the post-treatment weight" for the population is significantly different from 0.

[1 + 1 + 2 + 1 + 1 + 2 + 2 = 10]

6. (a) Model: $Y_i = \alpha + \beta x_i + e_i$ (or an equivalent form).

   Assumptions: Y_i is normally distributed; equal variance in the Y_is for each value of x_i (or $E_i \stackrel{d}{=} N(0, \sigma)$); observations are independent.

   Check normality with a normal probability plot of the residuals, which should be, roughly, a straight line.

   Check constant σ with a plot of the residuals versus the explanatory variable (or fitted values), which should be a random scatter about $y = 0$.

   (b) $\mathrm{H}_0 : \beta = 0, \qquad \mathrm{H}_1 : \beta \neq 0.$

   Either of the following tests:

   - $T = 5.21 \quad \Rightarrow \quad P\text{-value} < 0.001.$
   - $F = 27.12 \quad \Rightarrow \quad P\text{-value} < 0.001.$

   So there is evidence to reject the null hypothesis and we conclude that there is a linear relationship between stress level and distance travelled.

   (c) For zero distance travelled the stress level is 2.8. This may well be so, eg. work at home, but in that case the stress level has little or nothing to do with commuting distance.

   In the data set the distance travelled is in the range 4.5 to 18.4 km. Zero is outside this range and thus the fitted model may not apply. It is very risky to extrapolate.

   (d) $\frac{16.027}{21.937} \times 100 \approx 73.1\%$.

   (e) Adjusted $R^2 = \left[1 - \frac{MSE_{model}}{MSE_{total}}\right] \times 100 = \left[1 - \frac{0.591}{21.937/11}\right] \times 100 \approx 70.4\%$

   Adjusted R^2 adjusts the value of R^2, to take into account the estimation of 2 parameters in the linear model compared with the estimation of 1 parameter in the null model. The value of adjusted R^2 is less than the value of R^2.

   $[8 + 4 + 3 + 1 + 3 = 19$ marks$]$

7. (a) Significant interaction between size and species, so little point in examining main effects.

   Not much difference between species, except at size = large, where species A is substantially taller. (Lines diverge from parallelism at size = large.)

   Mean height increases for larger pots, except for the large pots of species B.

   (b) Need to compare means for size × species combinations: LSD $= 2.120 \times \sqrt{5.95(\frac{1}{3} + \frac{1}{3})} = 4.2$.

   Comparing largest two means for species A: $34.7 - 30.1 \pm 4.2 = (0.4, 9.8)$, which does not include 0, so recommendation would be: use size = large.

   Largest three means for species B do not differ significantly, so recommendation would be: use size = small, medium or large.

Species A:

Tube	Small	Med	Large
20.9	25.9	30.1	34.7

Species B:

Tube	Small	Large	Med
21.1	24.4	25.7	28.4

   (c) $\sqrt{5.95} = 2.44$.

   (d) Constant variance; satisfied—constant vertical scatter.

[4 + 5 + 2 + 3 = 14 marks]

8. (a) coefficient of x estimate (0.034) $\rightarrow$ z = 11.77 $\rightarrow$ P=0.000 (output: $< 2e^{-}16$);

   (b) residual deviance ($\chi^2_6 = 11.116$);

   (c) OR $= e^{0.034286} = 1.034881 = 1.035$;
   CI $= e^{0.034286 \pm 1.96 * 0.002913} = (1.028989, 1.040806) = (1.03, 1.04)$
   OR and CI interpretation: The odds of death of beetles exposed to pesticide increases by a factor of 1.035 compared to the odds of death for beetles not exposed to pesticide, and is between (1.03, 1.04) with 95% confidence.

   (d) $\hat{p} = \dfrac{e^{-5.88+250\times0.03}}{1+e^{-5.88+250\times0.03}} = 0.936$;

   (e) $-5.88 + 0.03 \times \text{LD50} = 0 \Rightarrow \text{LD50} = 171.6$.

[$1+1+2+2+2 = 8$ marks]

Total marks = 100

END OF EXAMINATION