Extended Answer Section

Answer these questions in the answer book(s) provided.
Ask for extra books if you need them.

1. Suppose the R objects `x` and `y` contain yields (in grams) of a certain crop from 12 plants from each of two varieties.

```
> sort(x)
 [1]  7.9  8.0  8.1  8.1  8.8  9.4  9.9 10.4 10.6 10.7 10.9 11.1
> sort(y)
 [1] 10.2 10.6 10.7 11.1 11.6 11.9 12.2 12.4 12.5 12.8 12.8 12.8
> mean(x)
[1] 9.491667
> mean(y)
[1] 11.8
> var(x)
[1] 1.579015
> var(y)
[1] 0.8872727
> var(x-y)
[1] 1.695379
```

(a) Perform an appropriate statistical test that makes normality and equal-variance assumptions to determine whether the mean difference between the groups is statistically significantly different from zero. Your answer should include

(i) a statistical model with appropriate parameters defined and formal hypotheses concerning the parameter(s);

(ii) a p-value, along with an interpretation of it.

(b) Sketch one or more boxplots and determine whether it/they validate or invalidate your assumptions.

2. (a) Ten plastic tiles each have a single letter printed on them as follows:

Letter	No. of tiles
S	3
T	3
I	2
A	1
C	1

If the tiles are randomly arranged in order, what is the probability that the word "STATISTICS" is obtained?

(b) A non-negative integer-valued random variable X satisfies $E\left(s^X\right) = \dfrac{(1+s)^5}{32}$ for all $s > 0$. Compute *Var(X)*.

(c) If X is Poisson(λ) for large λ then X is approximately normally distributed. Compute a normal approximation with continuity correction to $P(X \geq 32)$ when X is Poisson(25).

3. (a) Two random samples of size 200 are taken, one of female voters and another of male voters. Each is classified according to their preference between three political parties: the Tories, the Commies and the Hippies. The results are given in the table below:

	Tories	Commies	Hippies	Totals
Female	72	78	50	200
Male	90	70	40	200
Totals	162	148	90	400

(*i*) What kind of χ^2 test could be used to test the hypothesis that there is no real difference in preferences between female and male voters?

(*ii*) Obtain upper and lower bounds (the tightest possible using the tables) for the p-value for the test in the previous part. What is your conclusion?

(b) Suppose that a non-negative integer observation x is modelled as the value taken by a binomial(n, p) random variable X for n known but $0 \leq p \leq 1$ unknown. Writing $\hat{p} = x/n$ for the observed proportion, a Wilson confidence interval for p is the set of all p such that

$$-c \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq c$$

for a constant c.

(*i*) Show that the midpoint of the resultant interval is the following weighted average of $\hat{p}$ and $\frac{1}{2}$:

$$\left(\frac{n}{n+c^2}\right)\hat{p} + \left(\frac{c^2}{n+c^2}\right)\frac{1}{2}.$$

(*ii*) What value c should be used for the confidence level to be 99%?

NOT EXAMINABLE

4. (a) Suppose $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are ordered pairs, let a and b be the intercept and slope (respectively) of the corresponding least-squares line and for $i = 1, 2, \ldots, n$, write $\hat{\varepsilon}_i = y_i - (a + bx_i)$ for the i-th residual. Show that

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

where the quantities on the right-hand side have their conventional meanings as given in the formula sheet.

(b) The R objects `x` and `y` give (respectively) the sugar intake (grams per day) and weight (in kilograms) of 20 rugby league players. It is believed that there might be a positive linear association between sugar intake and bodyweight. It is thus decided to model the observed bodyweight values as random variables $Y_1, Y_2, \ldots, Y_{20}$ where for each $i = 1, 2, \ldots, 20$, $E(Y_i) = \alpha + \beta x_i$ for some unknown α and β.

Using the R output below compute an appropriate 99% confidence interval for the slope parameter β. Be careful to mention any further assumptions required for the interval to be valid.

```
> x

 [1] 13 14 24 25 22 23 20 18 19 36 11 21 19 12 15 28 10 22 16 24

> y

 [1]  95  87  94  88  84  84  99  95  91 101  74  85  94  87  88 115
[17]  84 102  82  96

> Sxx=sum((x-mean(x))^2)
> Syy=sum((y-mean(y))^2)
> Sxy=sum((x-mean(x))*(y-mean(y)))
> Sxx

[1] 768.8

> Syy

[1] 1537.75

> Sxy

[1] 643
```

End of Extended Answer Section