Useful formulae

Single group, n independent observations $x_1, \ldots, x_n$ from a population with mean μ and variance σ^2:

- sample mean $\hat{\mu} = \bar{x} = \sum x_i / n$
- sample variance $\widehat{\sigma^2} = s^2 = \sum (x_i - \bar{x})^2 \Big/ (n-1)$
- standard error of the sample mean is $\sigma / \sqrt{n}$

Two groups, n independent observations $x_1, \ldots, x_n$ from a population with mean μ_1 and m independent observations $y_1, \ldots, y_m$ from a population with mean μ_2 (common variance σ^2):

- standard error of mean difference is $\sigma \sqrt{(1/n) + (1/m)}$
- pooled estimate of common variance is $\left(\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2 \right) \Big/ (n + m - 2)$

1. (a) Given a random sample of n observations from a population distribution with mean μ and variance σ^2, derive the mean and variance of the sample mean.

[5 marks]

(b) A city planner asked a sample of 22 cyclists how long it took each of them to cycle home from work. The times, in minutes, were found to be as follows.

12	15	16	19	21	22	22	23	23	24	26
26	27	28	29	29	29	30	31	31	37	48

Form a stemplot and a boxplot of these data and describe the shape of the distribution.

[10 marks]

(c) Assume now that the data in part (b) come from a Normal distribution with mean 25 minutes and standard deviation 6 minutes. Find the probability that a randomly chosen cyclist from the population takes less than 20 minutes to cycle home.

What is the probability that the average journey time of a random sample of 10 cyclists from the population is less than 20 minutes?

[5 marks]

2. Define the *significance level* and *power* of a hypothesis test.

[2 marks]

An educational psychologist at a large university wants to estimate the mean IQ of the students in attendance. The IQs of a sample of 30 students are measured, and found to have sample mean 109.80, standard deviation 11.06.

(a) Assuming that IQ scores are Normally distributed, find a 95% confidence interval for the population mean IQ. Would the hypothesis that the mean IQ of students at the university is 100 be rejected in a two-sided test at the 5% significance level?

[5 marks]

(b) Suppose now that the standard deviation of IQ scores is known from previous studies to be 10.00, and does not have to be estimated from the sample. Find a 95% confidence interval for the population mean IQ, and comment on how the interval compares with that obtained in part (a).

[5 marks]

(c) Suppose another sample of 30 students is taken, and a one-sided test performed (using the known standard deviation of 10.00) to see whether mean IQ is greater than 100. Unknown to the experimenter, the true population mean is 105. Calculate the power of the test to detect this difference.

[8 marks]

3. A highway official wants to compare two brands of paint used for striping roads. Twenty locations are selected for paint stripes. The first brand is used on ten locations selected randomly from the twenty; the second brand on the remaining ten. The length of time each stripe lasts is recorded (in months), and summary statistics are given below.

	n	mean	s.d.
Brand A	10	36.12	0.950
Brand B	10	37.98	1.154

Assuming that the two samples are from Normal populations with common variance σ^2, obtain a pooled estimate of σ^2.

[2 marks]

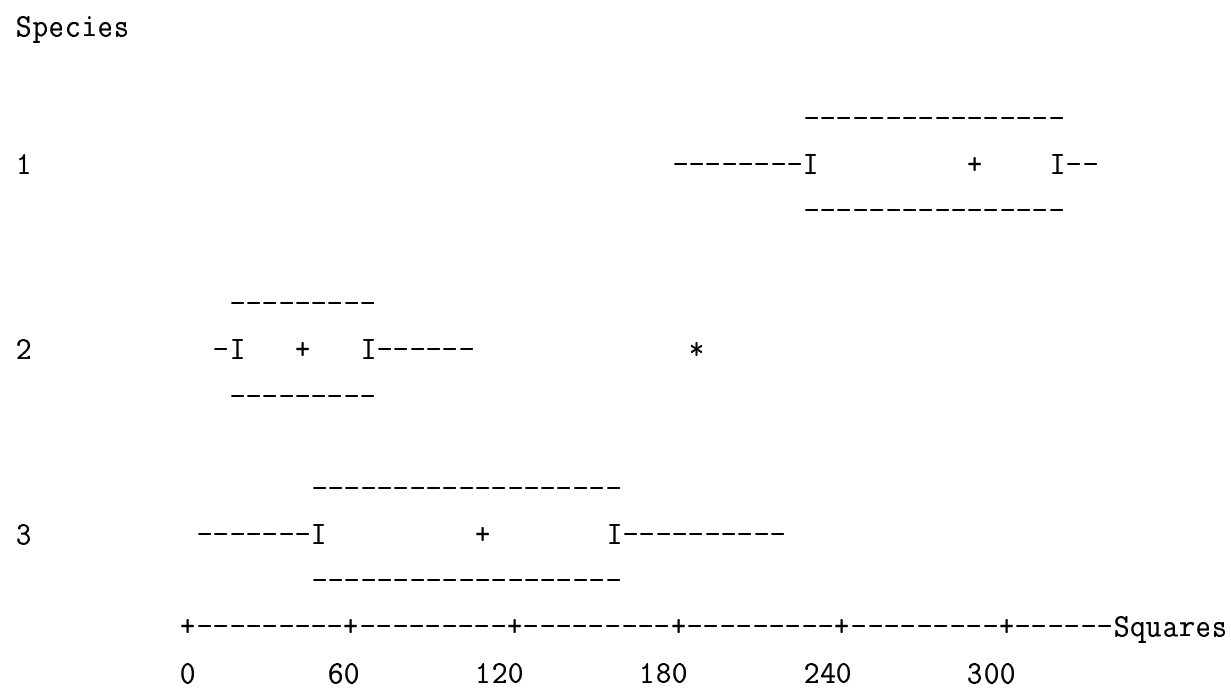
Test whether the population means for the two brands of paint differ.

[8 marks]

Compute a 90% confidence interval for the ratio of the true population standard deviations. Is it reasonable to assume a common variance?

[10 marks]

4. Forty-three mice of three different species were tested for 'aggressiveness' by placing each mouse in a box marked off into 49 equal squares and counting the number of squares traversed in a five-minute period. 16 mice of Species 1, 13 mice of Species 2, and 14 mice of Species 3 were tested in this way. Boxplots of the data are shown below.



Comment on the shape, spread and location of the data for the three species. Would performing a one-way analysis of variance on these data be justified? (Explain your answer.)

[6 marks]

The following Minitab output was obtained for a one-way analysis of variance.

Question 4 continued overleaf

One-Way Analysis of Variance

Analysis of Variance

Source	DF	SS	MS	F	P
Species	?	401978	?	?	?
Error	?	123801	?		
Total	?	?			

	N	Mean	StDev
Species 1	16	278.56	45.67
Species 2	13	54.54	49.54
Species 3	14	111.43	69.65

Test the hypothesis that there is no difference in aggressiveness between the three species.

[8 marks]

Compute a 95% confidence interval for the difference in means between Species 2 and Species 3.

[6 marks]

5. To investigate the effect of age (in years) upon heart rate during intensive exercise, ten randomly selected people performed exercise tests and recorded their peak heart rates. Analysis of the data using Minitab gave the following output.

```
Predictor       Coef       StDev          T        P
Constant     222.271       6.025      36.89    0.000
Age          -1.1405      0.1612      -7.07    0.000

S = 3.326       R-Sq = 86.2%     R-Sq(adj) = 84.5%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      553.60      553.60     50.04    0.000
Error         8       88.50       11.06
Total         9      642.10
```

Write down the estimated relationship between age and peak heart rate.

[1 mark]

Does age appear to have a significant effect upon heart rate, and does the variation in heart rates seem to be mostly due to age differences?

[2 marks]

Find 95% confidence intervals for the intercept and slope of the fitted line.

[6 marks]

Using x to denote `Age`, given that $\bar{x} = 36.80$ and $S_{xx} = \sum (x_i - \bar{x})^2 = 426.01$ find a 95% prediction interval for the peak heart rate of a 35 year old person.

[8 marks]

Explain the distinction between confidence intervals and prediction intervals.

[3 marks]

6. In multiple linear regression, explain how plots of standardised residuals may be used to check the validity of the model assumptions.

[6 marks]

Ten Corvettes were randomly selected from the classified ads in a newspaper, and for each car the age, miles driven and price were recorded. The data were read into Minitab and multiple regression performed, with explanatory variables

```
Age   = Age in years,
Miles = Thousands of miles driven,
```

and response variable

```
Price = Price in hundreds of dollars .
```

The following Minitab output was produced.

```
Regression Analysis

Predictor       Coef       StDev          T        P
Constant     287.362       9.943      28.90    0.000
Age          -37.375       5.174      -7.22    0.000
Miles         1.6378      0.8116       2.02    0.083

S = 12.11       R-Sq = 96.0%     R-Sq(adj) = 94.9%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    2       24655       12328     84.07    0.000
Error         7        1026         147
Total         9       25682
```

Write down the fitted model.

[1 mark]

Question 6 continued overleaf

For a 4 year old Corvette which has been driven 28000 miles, what would your prediction be for the price of the car?

[1 mark]

Interpret fully the given Minitab output.

[8 marks]

Suggest what further analysis you might carry out.

[4 marks]

7. What is a *contingency table*?

[2 marks]

A random sample of 307 lawyers were classified by type of practice and size of city in which they practice. The resulting data are shown below.

Type of practice	Size of city		
	Small	Medium	Large
Government	12	4	14
Judicial	8	1	2
Private practice	122	31	69
Salaried	19	7	18

Identify the *factors* in this example.

[1 mark]

For a lawyer chosen at random from those practicing in large cities, estimate the probability that they work in private practice.

[1 mark]

Work out expected frequencies under the hypothesis of no association between type of practice and size of city, and hence test the hypothesis of no association.

[11 marks]

Give a brief description of the data, taking into account the result of your hypothesis test, and comment on how the observed data compare with the expected frequencies under the hypothesis of no association.

[5 marks]