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) The following data are the failure times (in hundreds of hours) of 20 transmission mechanisms from caterpillar tractors belonging to a particular company.

44	40	26	23	12	33	69	40	32	59
24	75	39	95	45	22	13	51	22	17

Form a stemplot of these data and comment on the shape of the distribution. Find the value of the sample median.

Given that $\sum x_i = 781$ and $\sum (x_i - \bar{x})^2 = 9140.95$, obtain the values of the sample mean and sample standard deviation.

Would you prefer the sample median or sample mean as a measure of location for these data? Why?

(b) The annual precipitation in Illinois is approximately a Normal random variable with mean 84.6 cm and standard deviation 10.8 cm. What is the probability that in a particular year, the precipitation for the year lies between 90 cm and 100 cm?

What is the probability that the average annual precipitation over a given ten year period lies between 90 cm and 100 cm?

Find the level l such that the probability that the average annual precipitation over a given ten year period is less than l is 0.1.

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

For repairs to damaged cars, an insurance company can turn to two garages. To find which of the garages is cheaper, seven arbitrary cars are sent to both garage 1 and garage 2 to estimate the cost of repair. The table below gives the estimated cost (in hundreds of pounds) for each car and for each of the two garages.

Car	1	2	3	4	5	6	7
Garage 1	9.8	5.4	16.8	14.6	18.2	3.0	11.6
Garage 2	10.6	6.0	19.0	15.2	20.4	2.6	12.6

Assuming that the difference between the estimated cost of repairs to a car at garage 1 minus the estimated cost of repairs to the same car at garage 2 is approximately Normally distributed, obtain a 95% confidence interval for the mean difference in cost between the two garages. (The sample mean and sample standard deviation of the difference data are -1.0 and 0.931 respectively.)

Is the null hypothesis of zero mean difference between garages rejected at the 5% significance level in a two-sided test? What would you advise the insurance company?

If the standard deviation of the differences was known (rather than estimated) to be 0.9, how many observations would be needed for a one-sided test of the hypothesis of zero mean difference at the 5% significance level to have power 0.95 to detect a true mean difference of 0.5?

3. A statistician is asked to analyse the lengths of eggs from different species of birds. The experimenter supplies the sample mean lengths $\bar{x}_1 = 4.3$ cm, $\bar{x}_2 = 6.0$ cm from samples of ten eggs from each of two different species, together with the corresponding sample standard deviations $s_1 = 1.79$ cm, $s_2 = 2.00$ cm.

Assuming that the two samples of eggs are from populations with a common variance σ^2, obtain a pooled estimate of σ^2. Test whether the population mean lengths differ between the species.

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

4. A study was conducted to compare three types of rehydration solution in the treatment of acute diarrhoea in children. 30 children were divided at random into three groups of 10, each group treated with one of the rehydration solutions, and the weight gain (in grammes) during 24 hours recorded for each child.

	Weight gain									
Solution 1	40	44	119	133	138	157	204	208	219	420
Solution 2	21	49	60	67	88	92	100	112	121	148
Solution 3	-3	33	37	39	67	93	111	112	135	142

Draw boxplots of the data and comment.

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

```
One-Way Analysis of Variance


Analysis of Variance
Source          DF          SS        MS        F        P
Solution         ?       50883         ?        ?        ?
Error            ?      139724         ?
Total            ?           ?
```

Test the hypothesis that the solutions are equally effective. State clearly any assumptions that you make. Explain why the F statistic should give larger values under the alternative hypothesis that the solutions are not all equally effective than under the null hypothesis.

5. The following data are observations on the horsepower of an engine at 1800 rpm as a function of the viscosity of the oil, measured in suitable units.

Viscosity (x)	45	59	66	47	61	68	49	57	67	43
Horsepower (Y)	16.8	18.1	18.5	17.0	18.8	19.7	17.5	19.0	20.2	16.3

Suppose that the relationship between horsepower and viscosity can be modelled as

$$Y_i \quad = \quad \alpha + \beta x_i + \epsilon_i, \qquad (i = 1, 2, \ldots, 10)$$

where the ϵ_i are independent Normal random variables with mean zero and variance σ^2. Derive formulae for the least squares estimates of α and β, and give a formula for the estimation of σ^2.

Analysis of the data using Minitab gave the following output.

```
Predictor        Coef        StDev         T         P
Constant       11.174        1.013     11.03     0.000
Viscosit      0.12484      0.01779      7.02     0.000

S = 0.5093       R-Sq = 86.0%       R-Sq(adj) = 84.3%

Analysis of Variance
Source          DF         SS         MS         F         P
Regression       1     12.774     12.774     49.24     0.000
Error            8      2.075      0.259
Total            9     14.849
```

Write down the estimated relationship between horsepower and viscosity.

Given that $\bar{x} = 56.2$ and $\sum (x_i - \bar{x})^2 = 819.6$, find a 95% prediction interval for the horsepower of the engine when the viscosity of the oil is 60 units.

6. Define standardised residuals from fitting a linear regression model and describe a graphical method for testing the model assumption of Normality.

   In an investigation into ways of measuring the weight of plum trees without damaging them, nine mature trees were each pulled up and weighed. For each tree the circumference at the base of the trunk and the circumference at the top of the trunk were also recorded, the aim being to find a predictor of weight from these two circumference measurements. The data were read into Minitab, and regression performed after transformation of the variables, using explanatory variables

```
LNBASE = ln(circumference at the base of the trunk),
LNTOP  = ln(circumference at the top of the trunk),
```

   with response variable

```
LNWT   = ln(weight of tree).
```

   The following Minitab output was produced.

```
Regression Analysis

Predictor        Coef        StDev          T         P
Constant      -1.3209       0.6000      -2.20     0.070
LNBASE         1.2473       0.8372       1.49     0.187
LNTOP          0.8865       0.9072       0.98     0.366

R-Sq = 85.2%

Analysis of Variance

Source       DF          SS           MS         F        P
Regression    2    0.059068     0.029534     17.31    0.003
Error         6    0.010234     0.001706
Total         8    0.069302
```

Question 6 continued overleaf

Write down the fitted model. For a tree of base circumference 4.9 units and top circumference 4.8 units, what would be your prediction for the weight of the tree?

Interpret fully the given Minitab output, and suggest what further analysis you might carry out.

7. The table below shows the observed frequencies of different combinations of hair and eye colour for a group of 465 people.

	Hair colour		
Eye colour	Black	Brunette	Red
Brown	68	119	26
Blue	20	84	17
Hazel	15	54	14
Green	5	29	14

Work out the expected frequencies under the hypothesis of no association and comment on how these compare to the observed frequencies. Test the hypothesis of no association.