THE UNIVERSITY OF AUCKLAND

STATS 101
STATS 102
STATS 107
STATS 108

SECOND SEMESTER, 2001
Campus: CITY, TAMAKI

VERSION 1

STATISTICS

Introduction to Statistics
Statistics for Social Science
Statistics for Science & Technology
Statistics for Commerce

(Time allowed: THREE hours)

NOTE:

* **This examination consists of 65 multiple-choice questions.**
* **All questions have a single correct answer.**
* **If you give more than one answer to any question, you will receive zero marks for that question.**
* **No mark is deducted for an incorrect answer.**
* **All questions carry the same mark value.**
* **Answers must be written on the special answer sheet provided.**
* **Calculators are permitted.**

ATTACHMENT:

CONTINUED

Questions 1 to **5** refer to the following information.

Table 1 shows data for 10 NZ films funded by the New Zealand Film Commission (NZFC). The 10 films shown are the NZFC's top 10 with respect to **Percent Return**, which is calculated by expressing the film's total income from ticket sales (**Income**) as a percentage of the cost of production (**Budget**). For each film, measurements were made on the following variables:

Name:	Name of the film
Year:	Year of film's release
Budget:	Cost of production (× \$000's)
NZFC Invest:	Amount of NZFC funding (× \$000's)
Income:	Gross income (total box office receipts) (× \$000's)
Percent Return:	Percentage Return of Budget, ie **Income** expressed as a percentage of **Budget**

Name	Year	Budget	NZFC Invest	Income	Percent Return
Goodbye Pork Pie	1980	\$346	\$154	\$750	216%
Bad Taste	1986	\$295	\$207	\$1,034	350%
Never Say Die	1988	\$3,000	\$1,500	\$2,952	98%
An Angel At My Table	1990	\$2,789	\$2,789	\$5,095	183%
Braindead	1992	\$3,032	\$2,482	\$3,039	100%
Desperate Remedies	1993	\$2,100	\$1,600	\$1,300	62%
Once Were Warriors	1994	\$2,056	\$1,508	\$6,725	327%
Heavenly Creatures	1994	\$4,742	\$1,776	\$5,744	121%
The Ugly	1997	\$1,466	\$1,466	\$1,240	**
Topless Women Talk About Their Lives	1997	\$1,300	\$843	\$993	76%

Note:
1. \$ to the nearest thousand, ie \$346 means \$346,000.
2. ** denotes a missing value.

Table 1: Top 10 return of budget NZ films with NZFC funding (*NZFC figures*)

1. The main purpose of Table 1 is to compare **Percent Return** of the NZFC's 10 best films. One improvement in the presentation of the table would be to:

(1) present the measurements made on the three variables **Budget, Income** and **NZFC Invest** to the nearest dollar rather than the nearest thousand dollars.

(2) add a column to the right of the table giving row averages for each of the films.

(3) interchange the rows and the columns.

(4) express the value of **Percent Return** to at least one decimal place.

(5) list the films in order by the value of **Percent Return**.

2. In Table 1, the missing value, **, of **Percent Return** for the film 'The Ugly' is approximately:

(1) 121%

(2) 90%

(3) 100%

(4) 85%

(5) 118%

3. The sample mean, $\overline{x}$, and sample standard deviation, s, for the 10 values of **NZFC Invest** listed in Table 1 are:

(1) $\overline{x} = \$1,432,500 \quad s = \$855,238$

(2) $\overline{x} = \$1,432,500 \quad s = \$811,350$

(3) $\overline{x} = \$1,432,500 \quad s = \$731,432,500$

(4) $\overline{x} = \$2,887,200 \quad s = \$2,112,204$

(5) $\overline{x} = \$2,887,200 \quad s = \$2,226,459$

4. Figure 1 below shows the relationship between **NZFC Invest** and **Budget** for these 10 films.

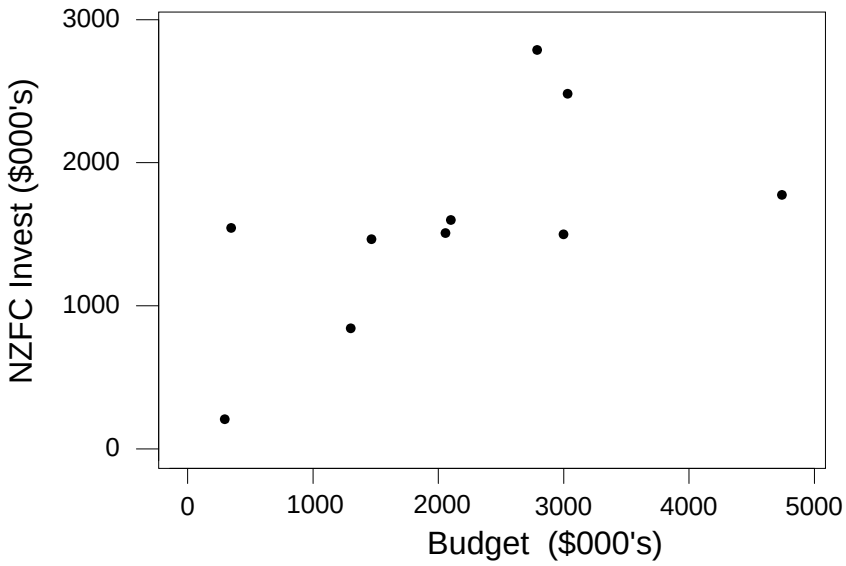

Figure 1: Scatter plot of **NZFC Invest** against **Budget**

Which **one** of the following statements is **true**?

(1) These data very clearly show that the relationship between these two variables is non-linear.

(2) The variables are on the correct axes, if we are using **Budget** as an explanatory variable and **NZFC Invest** as the response variable.

(3) There are two distinct groups in the **Budget** data.

(4) There is very clear constant scatter about a linear trend line.

(5) There is a negative association between the variables.

5. In 2000, the film *Crooked Earth* received \$2,336,700, the largest amount of funding awarded to a single film by the NZFC over the past 5 years. Figure 2 below shows dot plots of NZFC funding awarded to films in 1998 and 2000.

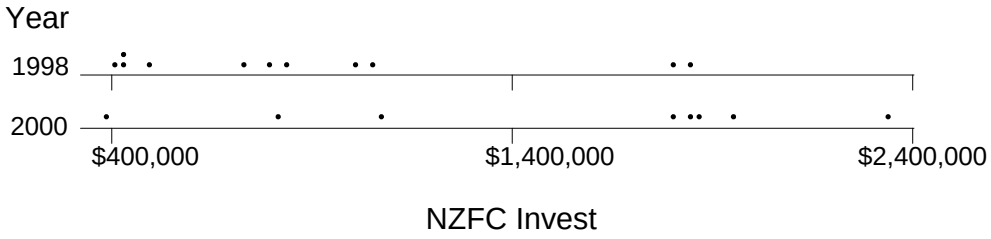

Figure 2: Dot plots of **NZFC Invest** in 1998 and 2000

Which **one** of the following statements is **true**?

(1) Two box plots would be preferred to compare the NZFC funding awarded in 1998 and 2000.

(2) The mean amount awarded in 2000 is smaller than the median amount awarded in that year.

(3) Labelled bar charts would be preferred to compare the NZFC funding awarded in 1998 and 2000.

(4) The standard deviation of the amount awarded in 1998 is higher than the standard deviation of the amount awarded in 2000.

(5) The average amount awarded in 1998 was higher than the average amount awarded in 2000.

6. Which **one** of the following statements is **false**?

(1) Plots can help determine which method(s) of analysis are appropriate.

(2) Plots are often the best tool to use first, when investigating the relationship between two quantitative variables.

(3) Outliers should always be removed before performing a formal analysis.

(4) Sample size may give some guidance as to what type of plot to use.

(5) Plots can reveal unusual or interesting features of the data.

Questions 7 to **10** refer to the following information.

The *NZ Herald*, 14 August 2001, reported the results of a two-year study at Hong Kong's Kwong Wah Hospital. The study comprised 5450 impotent men who were given Viagra at the hospital. Of these men, 3651 were smokers. Of the 926 impotent men for whom Viagra did **not** work, 840 were smokers.

7. The percentage of men in the study who were smokers and for whom Viagra did **not** work was approximately:

(1) 23.0%
(2) 51.6%
(3) 82.4%
(4) 90.7%
(5) 15.4%

8. The percentage of smokers for whom Viagra worked was approximately:

(1) 15.4%
(2) 77.0%
(3) 62.1%
(4) 51.6%
(5) 23.0%

Questions 9 and **10** refer to the following additional information.

Assume this sample of 5450 impotent men is a random sample of all impotent Chinese men and recall that Viagra did **not** work for 926 of the men in this sample. Suppose it is known that Viagra works for 78% of impotent Western males.

9. Which **one** of the following statements is **true**?

(1) The proportion of Chinese men in the sample for whom Viagra works is **both** an **estimate** value **and** a **parameter** value.

(2) The proportion of Chinese men in the sample for whom Viagra works is an **estimate** value whereas the '78%' of impotent Western males for whom Viagra works is a **parameter** value.

(3) The proportion of Chinese men in the sample for whom Viagra works and the '78%' of impotent Western males for whom Viagra works are **both parameter** values.

(4) The proportion of Chinese men in the sample for whom Viagra works and the '78%' of impotent Western males for whom Viagra works are **neither estimates nor parameter** values.

(5) The proportion of Chinese men in the sample for whom Viagra works is a **parameter** value whereas the '78%' of impotent Western males for whom Viagra works is an **estimate** value.

10. Suppose the results of this study are used to conduct a t-test to see if the percentage of impotent Chinese men for whom Viagra works has the same value as the percentage of impotent Western men for whom it is known to work. The formula for calculating the approximate standard error of the estimate in this test is:

(1) $\sqrt{\dfrac{0.67(1-0.67)}{5450} + \dfrac{0.78(1-0.78)}{5450}}$

(2) $\sqrt{\dfrac{0.83(1-0.83)}{5450} + \dfrac{0.78(1-0.78)}{5450}}$

(3) $\sqrt{\dfrac{(0.83+0.78) - (0.83-0.78)^2}{5450}}$

(4) $\sqrt{\dfrac{0.83(1-0.83)}{5450}}$

(5) $\sqrt{\dfrac{0.67(1-0.67)}{5450}}$

Questions 11 to **14** refer to the following information.

The owner of a pine forest knows that 8% of pine trees die before maturity. The forest is divided into 400 plots with 50 trees in each plot. Let X be the number of trees in a plot that die before maturity. Use a Binomial distribution with parameters $n = 50$ and $p = 0.08$ to model the distribution of the random variable X.

MINITAB output for the Binomial($n = 50$, $p = 0.08$) distribution is shown below.

```
Binomial with n = 50 and p = 0.08

x       Pr(X=x)      Pr(X<=x)
4        0.2037       0.6290
5        0.1629       0.7919
8        0.0271       0.9833
9        0.0110       0.9944
```

Table 2: MINITAB output: Binomial($n = 50$, $p = 0.08$)

11. The probability that, in a plot, 4 or 5 trees will die before maturity is approximately:

(1) 0.1629
(2) 0.2037
(3) 0.6290
(4) 0.3666
(5) 0.7919

12. The probability that, in a plot, at least 5 but less than 9 trees will die before maturity is approximately:

(1) 0.0000
(2) 0.1914
(3) 0.3654
(4) 0.2025
(5) 0.3543

13. Assume the 400 plots (each containing 50 trees) are independent of each other. The distribution of the mean number of trees (per plot) that die before maturity, $\overline{X}$, is best described as:

(1) approximately Binomial($n = 50$, $p = 0.08$)

(2) approximately Normal($\mu_{\overline{X}} = 4$, $\sigma_{\overline{X}} = 1.9183$)

(3) approximately Normal($\mu_{\overline{X}} = 4$, $\sigma_{\overline{X}} = 0.0959$)

(4) approximately Binomial($n = 400$, $p = 0.08$)

(5) approximately Normal($\mu_{\overline{X}} = 4$, $\sigma_{\overline{X}} = 0.0136$)

14. In this context, when using $X \sim$ Binomial($n = 50$, $p = 0.08$), which **one** of the following statements is **not** a necessary assumption?

(1) There are always the same number of trees per plot.

(2) The death before maturity of one pine tree does not result in the death before maturity of another pine tree.

(3) Two or more trees don't die before maturity at the same time.

(4) Environmental conditions affecting tree survival (soil quality, exposure to wind, sun etc) are the same for all trees.

(5) Each tree can be classified as either 'died before maturity' or 'survived to maturity'.

15. Let Y be a discrete random variable with mean $\mathrm{E}(Y) = 7$ and standard deviation $\mathrm{sd}(Y) = 3$. Let $W = 5 - 2Y$. The mean and standard deviation of W are:

(1) $\mathrm{E}(W) = -9, \quad \mathrm{sd}(W) = -6$

(2) $\mathrm{E}(W) = 19, \quad \mathrm{sd}(W) = 6$

(3) $\mathrm{E}(W) = 19, \quad \mathrm{sd}(W) = -6$

(4) $\mathrm{E}(W) = -9, \quad \mathrm{sd}(W) = -1$

(5) $\mathrm{E}(W) = -9, \quad \mathrm{sd}(W) = 6$

Questions 16 to **20** refer to the information given in **Appendix A: Racial Attitude Survey Data**, page 33.

16. An estimate of the difference between p_{more} and p_{less} is:

(1) 0.018
(2) −0.01
(3) −0.14
(4) 0.01
(5) −0.10

17. Information from Table 7, page 33, is used to construct a 95% confidence interval for the difference $p_{more} - p_{less}$. For the purpose of calculating $\text{se}(\widehat{p}_{more} - \widehat{p}_{less})$, the sampling situation can be described as:

(1) two independent samples of sizes 779 and 254.
(2) one sample of size 779, several response categories.
(3) one sample of size 1709, many yes/no items.
(4) one sample of size 1709, several response categories.
(5) one sample of size 779, many yes/no items.

18. A 95% confidence interval for the difference $p_{more} - p_{less}$ is $(-0.1833, -0.09668)$. The **best** interpretation of this interval is:

> With 95% confidence, the percentage of whites who feel that African Americans have more opportunities in life than whites have is somewhere between:

(1) 10% higher than and 18% lower than the percentage who feel that African Americans have less opportunities in life than whites have.
(2) 10% and 18%.
(3) 10% and 18% higher than the percentage who feel that African Americans have less opportunities in life than whites have.
(4) 10% lower than and 18% higher than the percentage who feel that African Americans have less opportunities in life than whites have.
(5) 10% and 18% lower than the percentage who feel that African Americans have less opportunities in life than whites have.

Questions 19 and **20** refer to the following additional information.

Let:

$p_{question1}$ be the proportion of Asian Americans who feel that **African Americans** have **more** opportunities in life than whites have

and

$p_{question2}$ be the proportion of Asian Americans who feel that **Asian Americans** have **more** opportunities in life than whites have.

Information from Table 7, page 33, is used to conduct a 2-tailed test for no difference between $p_{question1}$ and $p_{question2}$.

19. The sample proportions, $\widehat{p}_{question1}$ and $\widehat{p}_{question2}$, have come from *one sample, $n = 254$, many yes/no items*, ie Situation (c).

The formula for the standard error of the estimate, $\text{se}(\widehat{p}_{question1} - \widehat{p}_{question2})$, is:

(1) $\sqrt{\dfrac{(1-\widehat{p}_{question1}) + (1-\widehat{p}_{question2}) - (\widehat{p}_{question1} - \widehat{p}_{question2})^2}{254}}$

(2) $\sqrt{\dfrac{(\widehat{p}_{question1} + \widehat{p}_{question2}) - (\widehat{p}_{question1} + \widehat{p}_{question2})^2}{254}}$

(3) $\sqrt{\dfrac{\widehat{p}_{question1}(1-\widehat{p}_{question1})}{254} + \dfrac{\widehat{p}_{question2}(1-\widehat{p}_{question2})}{254}}$

(4) $\sqrt{\dfrac{\widehat{p}^2_{question1}}{254} - \dfrac{\widehat{p}^2_{question2}}{254}}$

(5) $\sqrt{\dfrac{(\widehat{p}_{question1} + \widehat{p}_{question2}) - (\widehat{p}_{question1} - \widehat{p}_{question2})^2}{254}}$

20. The expression for evaluating the test statistic for the null hypothesis, H_0: $p_{question1} - p_{question2} = 0$, is:

(1) $$\frac{\widehat{p}_{question1} - \widehat{p}_{question2}}{\text{se}(\widehat{p}_{question1}) + \text{se}(\widehat{p}_{question2})}$$

(2) $$\frac{\widehat{p}_{question1} - \widehat{p}_{question2}}{\sqrt{\text{se}(\widehat{p}_{question1})^2 - \text{se}(\widehat{p}_{question2})^2}}$$

(3) $$\frac{p_{question1} - p_{question2}}{\text{se}(\widehat{p}_{question1}) + \text{se}(\widehat{p}_{question2})}$$

(4) $$\frac{\widehat{p}_{question1} - \widehat{p}_{question2}}{\text{se}(\widehat{p}_{question1} - \widehat{p}_{question2})}$$

(5) $$\frac{p_{question1} - p_{question2}}{\text{se}(\widehat{p}_{question1} - \widehat{p}_{question2})}$$

21. Which **one** of the following statements is **false**?

(1) In hypothesis testing, statistical significance does not imply practical significance.

(2) In a hypothesis test for no difference between two means, a very small *P-value* indicates a very large difference in the means.

(3) In hypothesis testing, a non-significant test result does not imply that H_0 is true.

(4) In hypothesis testing, large samples can lead to small *P-values* without the results having any practical significance.

(5) In a hypothesis test for no difference between two means, a two-sided test should be used when the idea of doing the test has been triggered as a result of looking at the data.

Questions 22 and **23** refer to the following information.

Consider an investigation to compare the mean serum cholesterol levels (in mg/dl) produced by two diets. Some results from a two independent sample t-test are given below in Table 3.

$\overline{x}_1 - \overline{x}_2$	se$(\overline{x}_1 - \overline{x}_2)$	*P-value*	95% Confidence Interval
30mg/dl	16mg/dl	0.12	(−2mg/dl, 62mg/dl)

Table 3: Cholesterol study results

Note:

1. The hypotheses associated with the *P-value* given in the table are H_0: $\mu_1 - \mu_2 = 0$ and H_1: $\mu_1 - \mu_2 \neq 0$, where μ_1 and μ_2 are the underlying mean cholesterol levels produced by diet 1 and diet 2 respectively.
2. A difference in mean cholesterol level between the two diets of less than 10mg/dl is of no consequence, but a difference of 10mg/dl or more has important public health implications.

22. The test statistic for this 2-sample t-test has a value of approximately:

(1) 7.5
(2) 30
(3) 3.6
(4) 0.53
(5) 1.9

23. Which **one** of the following statements is **true**?

(1) The observed difference of 30mg/dl is statistically significant at the 5% level, but the difference in the underlying mean cholesterol levels is **not** large enough to be of practical significance (importance).

(2) The observed difference of 30mg/dl is **not** statistically significant at the 5% level, but the difference in the underlying mean cholesterol levels has been demonstrated to be large enough to be of practical significance (importance).

(3) The observed difference of 30mg/dl is **not** statistically significant at the 5% level, but if there really is a difference in the underlying mean cholesterol levels, it **could** be large enough to be of practical significance (importance).

(4) The observed difference of 30mg/dl is statistically significant at the 5% level and the difference in the underlying mean cholesterol levels is large enough to be of practical significance (importance).

(5) The observed difference of 30mg/dl is **not** statistically significant at the 5% level, but if there really is a difference in the underlying mean cholesterol levels, it would **not** be large enough to be of practical significance (importance).

Questions 24 and **25** refer to the following information.

It has been suggested that the risk of a car being stolen depends on the colour of the car. In a recent study 830 stolen cars were classified by colour. The results are shown in Table 4 below.

Colour	White	Red	Black	Other	**Total**
Count	271	134	231	194	**830**

Table 4: Car Colours

Suppose it is known that 35% of all cars are white, 15% are red, 30% are black and 20% are other colours. Assuming this sample is a random sample of all stolen cars, a Chi-square test is used to investigate the suggestion that the risk of a car being stolen depends on the colour of the car.

24. The correct hypotheses for this Chi-square test are:

(1) H_0: The proportions of stolen cars that are white, red, black and other colours are 0.35, 0.15, 0.30, and 0.20 respectively.

H_1: For stolen cars, at least one of the proportions stated in the null hypothesis is incorrect.

(2) H_0: For stolen cars, all of the proportions of each car colour are different.

H_1: For stolen cars, the proportion of each colour is the same.

(3) H_0: For stolen cars, the proportion of each car colour is the same.

H_1: For stolen cars, all of the proportions of each colour are different.

(4) H_0: For stolen cars, the proportion of each car colour is the same.

H_1: For stolen cars, at least one of the proportions of each colour is different from the rest.

(5) H_0: The proportions of stolen cars that are white, red, black and other colours are 0.35, 0.15, 0.30, and 0.20 respectively.

H_1: For stolen cars, none of the proportions stated in the null hypothesis is correct.

25. The correct expression for the Chi-square test statistic is:

(1) $\dfrac{(290.5-271)^2}{271}+\dfrac{(124.5-134)^2}{134}+\dfrac{(249-231)^2}{231}+\dfrac{(166-194)^2}{194}$

(2) $\dfrac{(271-290.5)^2}{271}+\dfrac{(134-124.5)^2}{134}+\dfrac{(231-249)^2}{231}+\dfrac{(194-166)^2}{194}$

(3) $\dfrac{(271-290.5)^2}{290.5}+\dfrac{(134-124.5)^2}{124.5}+\dfrac{(231-249)^2}{249}+\dfrac{(194-166)^2}{166}$

(4) $\dfrac{290.5-271}{271}+\dfrac{124.5-134}{134}+\dfrac{249-231}{231}+\dfrac{166-194}{194}$

(5) $\dfrac{271-290.5}{290.5}+\dfrac{134-124.5}{124.5}+\dfrac{231-249}{249}+\dfrac{194-166}{166}$

Questions 26 to **33** refer to the information given in **Appendix B: Class Survey Data**, pages 34 and 35.

26. The hypotheses for the Chi-square test described on page 34 are:

(1) H_0: The factors **Physical Activity** and **Gender** are not related.
H_1: The factors **Physical Activity** and **Gender** are related.

(2) H_0: The factors **Physical Activity** and **Gender** are not independent.
H_1: The factors **Physical Activity** and **Gender** are independent.

(3) H_0: The factors **Physical Activity** and **Gender** are not independent.
H_1: The factors **Physical Activity** and **Gender** are related.

(4) H_0: The factors **Physical Activity** and **Gender** are related.
H_1: The factors **Physical Activity** and **Gender** are not independent.

(5) H_0: The factors **Physical Activity** and **Gender** are independent.
H_1: The factors **Physical Activity** and **Gender** are not related.

27. Which **one** of the following statements is **false**?

(1) If the null hypothesis (H_0) is true, then the differences that we observe in the proportions $\frac{35}{45}$, $\frac{38}{67}$ and $\frac{4}{12}$ can be explained away simply as sampling variability.

(2) If the null hypothesis (H_0) is true, then the differences that we observe in the proportions $\frac{10}{45}$, $\frac{29}{67}$ and $\frac{8}{12}$ can be explained away simply as sampling variability.

(3) If the null hypothesis (H_0) is true, then the differences that we observe in the proportions $\frac{10}{47}$ and $\frac{35}{77}$ can be explained away simply as sampling variability.

(4) If the null hypothesis (H_0) is true, then the differences that we observe in the proportions $\frac{8}{47}$ and $\frac{4}{77}$ can be explained away simply as sampling variability.

(5) If the null hypothesis (H_0) is true, then the differences that we observe in the proportions $\frac{10}{45}$ and $\frac{38}{77}$ can be explained away simply as sampling variability.

28. The MINITAB output for this Chi-square test is shown in Table 5 below.

```
Expected counts are printed below observed counts
                  Physical Activity
Gender    Slight    Moderate    A Lot    Total
Male          10          29        8       47
           17.06       25.40     4.55

Female        35          38        4       77
           27.94       41.60     7.45

Total         45          67       12      124

Chi-Sq =  2.919   +  0.512  + 2.619  +
          1.782   +  0.312 +  1.599    = 9.743
DF = 2 ,  P-Value = 0.008
```

Table 5: MINITAB output

Which **one** of the following statements is **false**?

(1) One of the main reasons for such a small *P-value* in this test is the relatively large number of males who described their physical activity level as 'A Lot'.

(2) If the value of the test statistic had been 5.000 instead of 9.743, then the resulting *P-value* would have been larger than 0.008.

(3) There is very strong evidence of a relationship between the gender of an introductory statistics student and their description of their physical activity level.

(4) The validity of the test is doubtful since one of the cells has an expected count less than 5.

(5) One of the main reasons for such a small *P-value* in this test is the relatively small number of males who described their physical activity level as 'Slight'.

29. Using Table 9, page 35, an estimate of the difference between the underlying mean male and female pulse rate is approximately:

(1) -3.04 beats per minute

(2) -1.23 beats per minute

(3) 1.9 beats per minute

(4) -7.9 beats per minute

(5) 27 beats per minute

30. Using Table 9, page 35, the standard error of the difference between the two sample means, $\text{se}(\overline{x}_{Male} - \overline{x}_{Female})$, is approximately:

(1) -0.27
(2) 3.49
(3) 6.13
(4) 2.47
(5) 1.87

31. Which **one** of the following statements is **true**? (Use Table 9, page 35, to help answer this question.)

(1) If this analysis had been calculated by hand, then we would use $df = 44$ and the hand calculated confidence interval would be wider than that shown.
(2) The difference in the sample means is approximately 0.22 standard errors (of the difference) below zero.
(3) The probability that the null hypothesis, $\mu_{Male} = \mu_{Female}$, is true is approximately 0.22.
(4) If a one-tailed test was conducted on the data, then the resulting *P-value* would be 0.44.
(5) These data give evidence for a real difference between the underlying male and female mean pulse rate.

32. From Figures 3 and 4, page 35, the pulse rate of 23 beats per minute in the Female group was found to be a mistake — it was meant to have been 73 beats per minute. The observation is fixed and the two sample t-test is conducted again on the corrected data. Which **one** of the following statements is **true**?

(1) The new standard error of the difference in the sample means will be larger.
(2) The new confidence interval will be wider.
(3) The new *P-value* will be smaller.
(4) The new difference between the sample means will be smaller.
(5) The new test statistic will be closer to zero.

33. Which **one** of the following statements is **true**?

(1) The variable **Gender** is a quantitative variable.
(2) The variable **Pulse** is a categorical qualitative variable.
(3) A scatter plot of **Physical Activity** against **Gender** is an appropriate tool to explore the relationship between these two variables.
(4) The variable **Pulse** can be treated as a continuous random variable because it has a relatively small number of repeated values.
(5) The variable **Physical Activity** is a discrete random variable.

Questions 34 to **39** refer to the information given in **Appendix C: Marijuana Survey Data**, page 36.

34. The null hypothesis, H_0, for the test described on page 36 is:

(1) H_0: The years 1990 and 1998 are independent.

(2) H_0: The proportion of males is the same for each level of the factor **Age** in both 1990 and 1998.

(3) H_0: The factors **Age** and **Year** are not independent.

(4) H_0: For a given age group, the proportion of males who have used marijuana in the last 12 months is the same in 1990 as in 1998.

(5) H_0: The distribution of **Age** is different for each level of the factor **Year**.

35. The degrees of freedom, df, for the Chi-square test shown in Table 11, page 36, has a value:

(1) $df = 8$

(2) $df = 5$

(3) $df = 14$

(4) $df = 6$

(5) $df = 7$

36. Consider the cell in Table 11, page 36, for **1998** and 25–29 year-old males. The expected cell count, under the null hypothesis, for 25–29 year-old males in **1998** is:

(1) 366

(2) 127.75

(3) 243

(4) 122

(5) 115.25

37. Consider the cell in Table 11, page 36, for **1990** and 25–29 year-old males. This cell's contribution to the Chi-square test statistic value of 7.736 is approximately:

(1) 0.0025

(2) 0.2732

(3) −0.0475

(4) 0.0499

(5) 0.2869

38. The *P-value* for this test is 0.258. Which **one** of the following conclusions is **true**?

(1) There is no difference between 1990 and 1998 for the age distributions of males who have used marijuana in the last 12 months.

(2) For a given age group, the proportion of males who have used marijuana in the last 12 months is the same in 1990 and 1998.

(3) The data is consistent with having come from 2 populations which have the same age distributions for males who have used marijuana in the last 12 months.

(4) If the age distribution of males who have used marijuana over the last 12 months in 1990 and 1998 were the same, then it would have been very unusual to have obtained data like this.

(5) There is evidence that the 1990 and 1998 age distributions of males who have used marijuana in the past 12 months are not the same.

39. Suppose it had been decided in 1998 to have used a sample size of approximately 2500 people, approximately four times the size of the sample in 1990. The effect of this increase would have made:

(1) no difference to the precision of the 1998 estimates since the sample size would have still been less than 10% of the population size.

(2) the margin of errors in the 1998 estimates approximately one quarter the size of those in 1990.

(3) the variabilities of the 1998 estimates smaller than those of the 1990 estimates.

(4) the biases in the 1998 estimates larger than those in the 1990 estimates.

(5) the standard errors of the 1998 estimates larger than the corresponding standard errors in 1990.

40. An experiment was designed to investigate the effect of the amount of water and seed variety upon subsequent growth of plants. Each plant was randomly allocated to, and potted in, a clay pot. Each clay pot (and plant) was then randomly allocated a measured amount of water to be given weekly. The height of the plant was measured at the end of the experiment. Which **one** of the following statements is **false**?

(1) The response variable is height.

(2) The explanatory variables are seed variety and amount of water.

(3) The use of randomisation ensured that the effects of other possible factors were eliminated.

(4) A possible uncontrollable factor in this experiment is any nutrients that might have been present in the pot.

(5) Designed experiments like this one, give the best evidence of "cause and effect" relationships.

41. Which **one** of the following statements is **true**?

A nonparametric equivalent of a two-sample t-test is a:

(1) Mann-Whitney (Wilcoxon rank-sum) test.

(2) Chi-square test.

(3) Kruskal-Wallis test.

(4) Welch test.

(5) one-way analysis of variance F-test.

42. Given a random sample from a population with a population proportion p, which **one** of the following statements is **false**?

(1) The distribution of the sample proportion, $\widehat{P}$, is exactly Normally distributed if the distribution of the population is exactly Normal.

(2) The sample proportion is an unbiased estimate of the population proportion since $\mathrm{E}(\widehat{P}) = p$.

(3) The standard error of the sample proportion is an estimate of the standard deviation of the sample proportion.

(4) The mean of the distribution of the sample proportion is equal to the population proportion. Ie, $\mu_{\widehat{P}} = p$.

(5) For large samples, the distribution of the sample proportion, $\widehat{P}$, is approximately Normally distributed.

Questions 43 to **57** refer to the information given in **Appendix D: Radiata Pine Tree Data**, pages 37 to 43.

Questions 43 and **44** refer to the following additional information.

An attempt is made to use the variable **Density** to explain the behaviour of the variable **Diameter**. The results of a linear regression analysis and associated plots are shown in Figure 5 and Table 13, (page 38), and Figures 6 to 8 (page 39).

43. Under this linear regression model, which of the following statements are believable? (Use the figures and table on pages 38 and 39 to answer this question.)

I:	The standard deviation of the errors is not related to the **Density** value.
II:	There is no linear relationship between the variables.
III:	The errors are Normally distributed.
IV:	The mean of the errors is zero.
V:	The errors are independent with respect to observation order.

(1) All of the statements **I**–**V**

(2) **I and III** only

(3) **II, III** and **IV** only

(4) **III** only

(5) **IV** only

44. The *P-value* at the end of the first row of Table 13 is related to a significance test. Which **one** of the following is the correct null hypothesis, H_0, for this test?

(1) H_0: The intercept is zero.

(2) H_0: The intercept is 38.421.

(3) H_0: The slope is 38.421.

(4) H_0: The estimate of the slope is zero.

(5) H_0: The estimate of the intercept is zero.

Questions 45 to **52** refer to the following additional information.

A simple linear regression is conducted with **Density** as the explanatory variable and **TrunkScore** as the response variable. The regression analysis output and associated plots are given in Figure 9 and Table 14 (page 40), and Figures 10 to 12 (page 41).

45. **Using only Figure 10**, page 41, a residual scatter plot of **Residual** against **Density**, we can conclude that:

(1) the mean of the errors is not zero.

(2) the standard deviation of the errors is not constant for all **Density** values.

(3) there is no indication of a violation of any of the linear regression assumptions.

(4) there is a non-linear relationship between **Density** and **TrunkScore**.

(5) **Density** is depicted as a continuous quantitative variable.

46. The expected value of **TrunkScore** for a randomly selected tree from a one hectare sized plot containing 2000 trees is approximately: (Use Table 14 (page 40) to help answer this question.)

(1) 11.4

(2) 4.5

(3) 4.6

(4) 4.7

(5) 12.6

47. One of the trees in the sample had a **TrunkScore** value of 4.171 and a **Density** value of 2.425. Based on the regression equation,

$$\text{TrunkScore} = 6.90 - 1.20\ \text{Density}$$

the residual for this tree is approximately:

(1) 0.18

(2) −0.18

(3) 1.23

(4) 3.99

(5) 4.17

48. The regression equation

$$\text{TrunkScore} = 6.90 - 1.20 \text{ Density}$$

predicts that:

(1) the average trunk score of a tree decreases by 0.015 for every extra 1000 trees per hectare.

(2) the trunk score of a tree increases by 1.20 for every extra 1000 trees per hectare.

(3) the average trunk score of a tree increases by 6.90 for every extra 1000 trees per hectare.

(4) the average trunk score of a tree decreases by 6.90 for every extra 1000 trees per hectare.

(5) the average trunk score of a tree decreases by 1.20 for every extra 1000 trees per hectare.

49. In a test for no relationship between **TrunkScore** and **Density**, the appropriate hypotheses are:

(1) $H_0 : \widehat{\beta}_1 \neq 0 \qquad H_1 : \widehat{\beta}_1 = 0$

(2) $H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$

(3) $H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 > 0$

(4) $H_0 : \beta_1 \neq 0 \qquad H_1 : \beta_1 = 0$

(5) $H_0 : \widehat{\beta}_1 = 0 \qquad H_1 : \widehat{\beta}_1 \neq 0$

50. The degrees of freedom associated with the test in Question **49** are:

(1) 873

(2) 869

(3) 871

(4) 872

(5) 870

51. Two one-hectare sized plots differ in that one of them has 875 more trees than the other. The difference expected in the trunk scores between trees from the two plots is approximately:

(1) 1.05

(2) 5.85

(3) 1

(4) 2.95

(5) 4.84

52. Which **one** of the following statements about this simple linear regresssion is **false**?

(1) The fitted line, used to summarise the relationship between the two variables, is chosen to maximise the overall size of the residuals.

(2) It is unwise to make predictions outside the range of the explanatory variable, **Density**.

(3) We use simple linear regression to study the relationship between these two quantitative variables.

(4) We model the relationship between the two variables as consisting of "trend" plus "scatter".

(5) The response variable, **TrunkScore**, in this simple linear regression model must be a random variable.

Questions 53 to **57** refer to the following additional information.

The size of a tree may also be affected by the altitude at which the tree is planted. Altitude was classified as either Low, Medium or High. Page 42 shows Descriptive Statistics (Table 15), dot plots (Figure 13), and box plots (Figure 14) for diameters of trees at each of the three different altitude groups.

53. Suppose that the three altitude samples (Low, Medium, and High) are independent. We are considering the appropriateness of using a one-way analysis of variance (ANOVA) F-test to test for no difference in the underlying mean diameters of trees at different altitude levels. Which **one** of the following statements is **true**? (Use Table 15 and Figures 13 and 14, page 42, to help answer this question.)

(1) It is **not** appropriate to use the F-test because the F-test is **not** sufficiently robust to withstand the departures from Normality suggested by the samples.

(2) It is **not** appropriate to use the F-test, because the F-test is **not** sufficiently robust to withstand the departures from the assumption of equal group population standard deviations suggested by the samples.

(3) It is appropriate to use the F-test because the F-test is sufficiently robust to withstand both the departures from Normality and the departures from the assumption of equal group population standard deviations suggested by the samples.

(4) It is **not** appropriate to use the F-test because there are only three samples.

(5) It is **not** appropriate to use the F-test because two of the box plots show outside values.

For **Questions 54** to **57** assume that the use of the F-test is appropriate.

54. A one-way analysis of variance (ANOVA) F-test is conducted on the data from this sample of 871 trees to investigate whether altitude has an effect on the size of the diameter of a tree. The output from the analysis is given in Table 16, page 43. The values for the degrees of freedom, `df1` and `df2`, for this F-test are:

(1) `df1` $= 2$, `df2` $= 868$

(2) `df1` $= 2$, `df2` $= 869$

(3) `df1` $= 2$, `df2` $= 871$

(4) `df1` $= 3$, `df2` $= 869$

(5) `df1` $= 3$, `df2` $= 868$

55. The value of the F-test statistic, `fo`, for the F-test refered to in Question **54** is approximately:

(1) 2.000

(2) 0.2536

(3) 126.4

(4) 0.5000

(5) 8963

56. Which **one** of the following is the **best** interpretation of the results of this F-test?

(1) There is very strong evidence that the sample means for the three altitude levels are not all the same.

(2) There is very strong evidence against the hypothesis that the underlying means for the three altitude levels are all different.

(3) There is very strong evidence against the hypothesis that the underlying means for the three altitude levels are not all the same.

(4) The results are inconclusive since the *P-value* has a value of 0.000.

(5) There is very strong evidence that the underlying means for the three altitude levels are not all the same.

57. Which **one** of the following statements about the pairwise comparisons in Table 16, page 43, is **true**?

(1) With 95% confidence, the underlying mean diameter of **high** altitude trees is somewhere between 6.932cm and 10.053cm greater than the underlying mean diameter of **low** altitude trees.

(2) The difference between the observed sample means for the group of **high** altitude trees and the group of **medium** altitude trees is significant, at the 5% level of significance.

(3) With 95% confidence, the underlying mean diameter of **medium** altitude trees is somewhere between 1.765cm less than and 1.560cm greater than the underlying mean diameter of **low** altitude trees.

(4) In a two-tailed test for no difference between the underlying mean diameter of **medium** altitude trees and the underlying mean diameter of **low** altitude trees, the *P-value* will be less than 1%.

(5) The difference between the observed sample means for the group of **medium** altitude trees and the group of **low** altitude trees is significant, at the 5% level of significance.

58. Which **one** of the following statements is **false**:

(1) A strong linear relationship between two quantitative variables is indicated when a scatter plot of sample data shows a clear linear trend with very little scatter, giving rise to a sample correlation coefficient close to 1 or -1.

(2) The least squares method chooses the line which minimises the overall size of the residuals (as measured by the sum of the squared prediction errors).

(3) A single outlier can exert a strong influence over the least squares line.

(4) A linear regression model should never be used without first examining the appropriate scatter plot.

(5) A correlation coefficient of zero between sample data on two variables implies that the variables are completely unrelated.

59. An experiment, *Get Going With Breakfast*, was recently conducted to demonstrate the importance of eating breakfast to teenagers. One hundred and sixty pupils, from five different secondary schools throughout New Zealand, were given breakfast every day for the month of August. Changes in their energy levels and physical and mental performance at school were monitored. The classes were also involved in special mentoring sessions each week with some of New Zealand's top sports people.

Which **one** of the following statements is **NOT** a potential problem with this experiment?

(1) The sample size is too small compared with the total number of New Zealand teenagers.

(2) The effect of the daily breakfast on academic performance may be confounded with the effect of the 'special mentoring sessions'.

(3) There is no mention of the use of a control group.

(4) There is no mention that some form of randomisation was used.

(5) A possible placebo effect was ignored.

Questions 60 to **63** refer to the information given in **Appendix E: Fungi Data**, page 44.

60. The interquartile range for X_H is:

(1) 1.6cm
(2) 2.4cm
(3) 3.0cm
(4) 2.5cm
(5) 2.1cm

61. Under the suggested classification plan, the probability of classifying a fungus as poisonous *Sulphur Tuft* when it really is an edible *Hypholoma Capnoides* is approximately:

(1) 0.80
(2) 0.25
(3) 0.88
(4) 0.20
(5) 0.12

62. To ensure that the probability of mistaking a poisonous *Sulphur Tuft* for an edible *Hypholoma Capnoides* is 1%, the threshold height should be changed from 8cm to approximately:

(1) 6.5cm
(2) 12.4cm
(3) 2.4cm
(4) 6.6cm
(5) 10.6cm

63. If we want to find the probability that a randomly chosen *Sulphur Tuft* fungus and a randomly chosen *Hypholoma Capnoides* fungus differ in height by no more than 1cm, then we need to compute:

(1) $1 - \text{pr}(-1 \leq X_S - X_H \leq 1)$ where $X_S - X_H \sim \text{Normal}(\mu = 3,\ \sigma = 3.01)$
(2) $\text{pr}(X_S - X_H \leq 1)$ where $X_S - X_H \sim \text{Normal}(\mu = 3,\ \sigma = 3.01)$
(3) $1 - \text{pr}(-1 \leq X_S - X_H \leq 1)$ where $X_S - X_H \sim \text{Normal}(\mu = -3,\ \sigma = 2.16)$
(4) $\text{pr}(-1 \leq X_S - X_H \leq 1)$ where $X_S - X_H \sim \text{Normal}(\mu = 3,\ \sigma = 2.16)$
(5) $\text{pr}(X_S - X_H \leq 1)$ where $X_S - X_H \sim \text{Normal}(\mu = 3,\ \sigma = 2.16)$

Questions 64 and **65** refer to the following information.

An insurance company wanted to compare the repair costs for cars at two particular garages. Repair cost estimates obtained on the same 8 cars from the two garages are given in Table 6 below.

	Car							
Garage	1	2	3	4	5	6	7	8
Garage 1	\$700	\$1500	\$2000	\$850	\$1450	\$2300	\$1800	\$2000
Garage 2	\$800	\$1300	\$2000	\$800	\$1100	\$2500	\$1650	\$1950

Table 6: Repair cost estimates of the same eight cars at two garages

64. A 2-sided sign test is used to test the hypothesis that, on average, the garages' repair cost estimates are the same. Which **one** of the following statements is **false**?

(1) It is necessary to assume that the sample of repair cost estimates from Garage 1 is independent of the sample from Garage 2.

(2) A one sample t-test would be the equivalent parametric test to use in this situation.

(3) For each car, it is necessary to consider the difference in repair cost estimates between the garages.

(4) It is not necessary to assume that the underlying distributions of repair costs for each of the garages are Normally distributed.

(5) For each garage, it is necessary to assume that the 8 repair cost estimates are independent.

65. The *P-value* for the 2-sided sign test can be calculated by:

(1) $2 \times \text{pr}(Y \geq 5)$ where $Y \sim \text{Binomial}(n = 8, p = 0.5)$

(2) $2 \times \text{pr}(Y \geq 4)$ where $Y \sim \text{Binomial}(n = 8, p = 0.5)$

(3) $2 \times \text{pr}(Y \leq 4)$ where $Y \sim \text{Binomial}(n = 8, p = 0.5)$

(4) $2 \times \text{pr}(Y \geq 5)$ where $Y \sim \text{Binomial}(n = 7, p = 0.5)$

(5) $2 \times \text{pr}(Y \geq 4)$ where $Y \sim \text{Binomial}(n = 7, p = 0.5)$

INCLUSIONS:

* **Appendix A: Racial Attitude Survey Data** for use in Questions **16 to 20**
* **Appendix B: Class Survey Data** for use in Questions **26 to 32**
* **Appendix C: Marijuana Survey Data** for use in Questions **34 to 38**
* **Appendix D: Radiata Pine Tree Data** for use in Questions **43 to 57**
* **Appendix E: Fungi Data** for use in Questions **60 to 63**
* **Formulae Appendix**

BLANK PAGE

Appendix A: Racial Attitude Survey Data

Questions 16 to **20** refer to the information given in this appendix

The Washington Post, The Henry J Kaiser Family Foundation and *Harvard University* conducted a poll (8 March – 22 April, 2001) 'to gauge the racial attitudes of American adults'. The telephone poll surveyed 1709 adults including 779 whites, 323 African Americans, 315 Hispanics and 254 Asian Americans. Assume this sample of 1709 adults is a random sample of American adults. Two of the questions in the survey were:

> **Question 1:**
> Do you feel that **African Americans** have more, less or about the same opportunities in life as whites have?

and

> **Question 2:**
> Do you feel that **Asian Americans** have more, less or about the same opportunities in life as whites have?

The percentage results for these two questions are shown in Table 7 below.

	Response				Sample size
	More %	Less %	Same %	Unsure %	
Question 1					
White	13	27	58	2	779
African American	1	74	23	2	323
Hispanic	8	46	44	2	315
Asian American	10	44	39	7	254
Total Sample	11	35	51	2	1709
Question 2					
White	13	14	70	4	779
African American	15	38	39	8	323
Hispanic	18	24	55	3	315
Asian American	7	34	53	5	254
Total Sample	14	18	63	4	1709

Table 7: Americans' responses to racial attitudes survey

Let p_{more} be the proportion of **whites** who feel that African Americans have **more** opportunities in life than whites have and p_{less} be the proportion of **whites** who feel that African Americans have **less** opportunities in life than whites have. (Ie, whites' responses to Question 1.)

Appendix B: Class Survey Data

Questions 26 to **33** refer to the information given in this appendix.

A class survey is conducted in the first lecture of introductory statistics courses. One question asks students to describe the level of their physical activity as: slight, moderate or a lot. The results of this question for 124 students from one stream of STATS 108 were cross classified by gender and are shown in Table 8 below.

	Physical Activity			
Gender	Slight	Moderate	A Lot	**Totals**
Male	10	29	8	47
Female	35	38	4	77
Totals	45	67	12	124

Table 8: Self described physical activity level for introductory statistics students

Assume these 124 students are a random sample of all introductory statistics students. A Chi-square test is conducted to investigate the relationship between **Gender** and **Physical Activity**.

Another question asked students to take their pulse and record their pulse rate in beats per minute. 117 students responded. Dot plots and box plots of **Pulse** by **Gender** are shown in Figures 3 and 4. A two sample t-test is conducted on the data using MINITAB and the output is shown in Table 9.

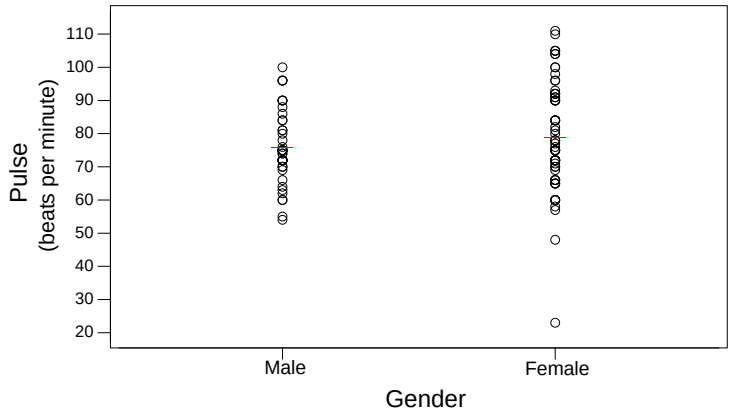

Figure 3: Dot plot of **Pulse** against **Gender**

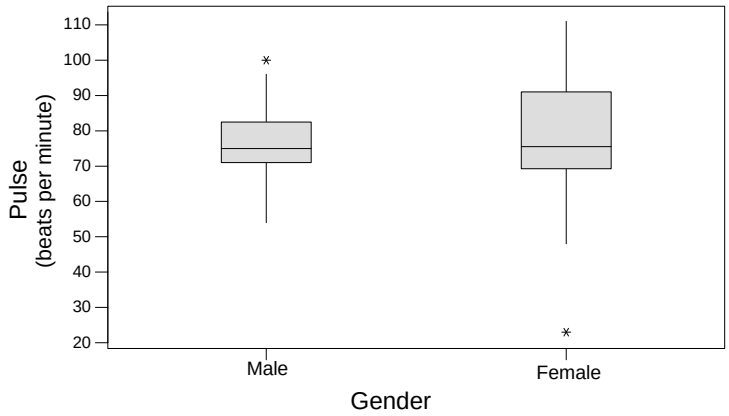

Figure 4: Box plot of **Pulse** against **Gender**

Two Sample T-Test and Confidence Interval

```
Two sample T for Pulse

Gender          N       Mean     StDev    SE Mean
Male           45      75.78     10.80       1.61
Female         72      78.82     15.94       1.88

95% CI for mu (Male) - mu (Female): (-7.9,  1.9)
T-Test mu (Male) = mu (Female) (vs not =): T = -1.23  P = 0.22  DF = 114
```

Table 9: MINITAB output for difference between population means for **Pulse**

Appendix C: Marijuana Survey Data

Questions 34 to **39** refer to the information given in this appendix.

The article 'Marijuana use in New Zealand, 1990 and 1998' (*New Zealand Medical Journal, 10 August 2001*) examines changes in marijuana use in an urban region and a rural region in New Zealand between 1990 and 1998. Two random sample surveys of people aged 15–45 years were carried out in Auckland and Bay of Plenty regions in 1990 and 1998, using a computer-assisted telephone interviewing system. Table 10 shows the age distribution of the number of males reported using marijuana in the last 12 months for 1990 and 1998.

	Age							
Year	15–17	18–19	20–24	25–29	30–34	35–39	40–45	**Totals**
1990	70	82	159	121	91	59	27	609
1998	95	90	160	122	88	73	47	675

Table 10: Number of males who reported using marijuana in the last 12 months

We used MINITAB to conduct a Chi-square test to investigate any differences between 1990 and 1998 for the age distribution of male marijuana users. The output is given in Table 11 below. Some values have been removed and replaced with an asterisk (∗).

```
Expected counts are printed below observed counts
                                   AGE
Year       15-17   18-19   20-24   25-29   30-34   35-39   40-45   Total

1990          70      82     159     121      91      59      27     609
           78.26   81.58  151.30  115.25   84.90   62.61   35.10

1998          95      90     160     122      88      73      47     675
              *   90.42  167.70       *   94.10   69.39   38.90

Total        165     172     319     243     179     132      74    1284

Chi-Sq = 7.736
DF = * ,  P-Value = 0.258
```

Table 11: MINITAB output

Appendix D: Radiata Pine Tree Data

Questions 43 to **57** refer to the information given in this appendix.

In their effort to maximise profits, forest owners are interested in the factors which have an effect on the size of a tree. A random sample of 871 Radiata Pine trees planted in 1971 was taken from plots in the Bay of Plenty region. For each tree, measurements were made on the following variables:

Diameter: Tree diameter (cm) at 1.4m above the ground level.
TrunkScore: A trunk size score related to a measure of the trunk diameter.
Density: The number of thousands of trees per hectare for the plot.
Altitude: Altitude of the plot — High, Medium or Low.

A section of the data collected is shown in Table 12 below.

No.	Diameter	TrunkScore	Density	Altitude
1	15.0	2.86437	3.150	High
2	9.8	3.01664	3.150	High
3	12.0	3.02700	3.150	High
4	17.4	3.51665	3.150	High
5	19.5	3.54948	3.075	Med
⋮	⋮	⋮	⋮	⋮
868	20.5	4.22247	2.175	Med
869	21.9	4.15398	2.000	Low
870	22.8	4.6357	1.925	Med
871	23.6	4.99768	1.650	Med

Table 12: Radata Pine data

Diameter versus Density

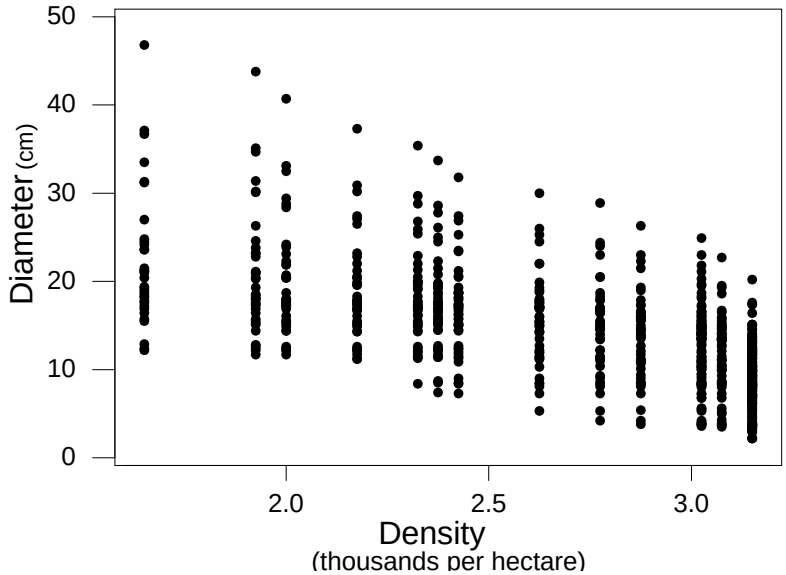

Figure 5: Scatter plot of **Diameter** against **Density**

Regression Analysis for Diameter on Density

```
The regression equation is
Diameter = 38.4 - 8.97 Density

Predictor       Coef      StDev          T          P
Constant      38.421      1.023      37.57      0.000
Density       -8.971      0.369     -24.30      0.000

S = 4.912      R-Sq = 40.5%    R-Sq(adj) = 40.4%
```

Table 13: Regression analysis of the relationship between **Diameter** and **Density**

Some model checking plots are shown on the next page.

Response is Diameter

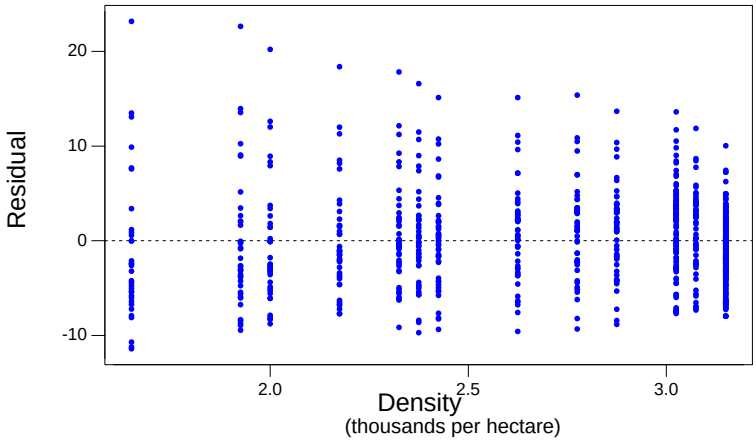

Figure 6: Residual plot of **Residual** against **Density**

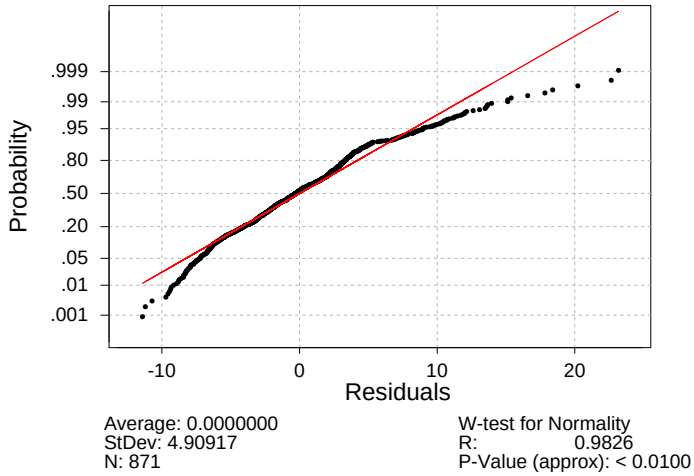

Figure 7: Normal probability plot of **Residual**

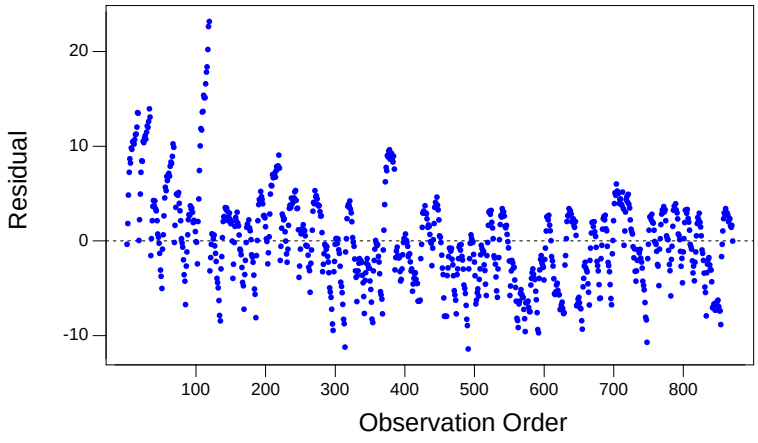

Figure 8: Scatter plot of **Residual** against **Observation Order**

TrunkScore versus Density

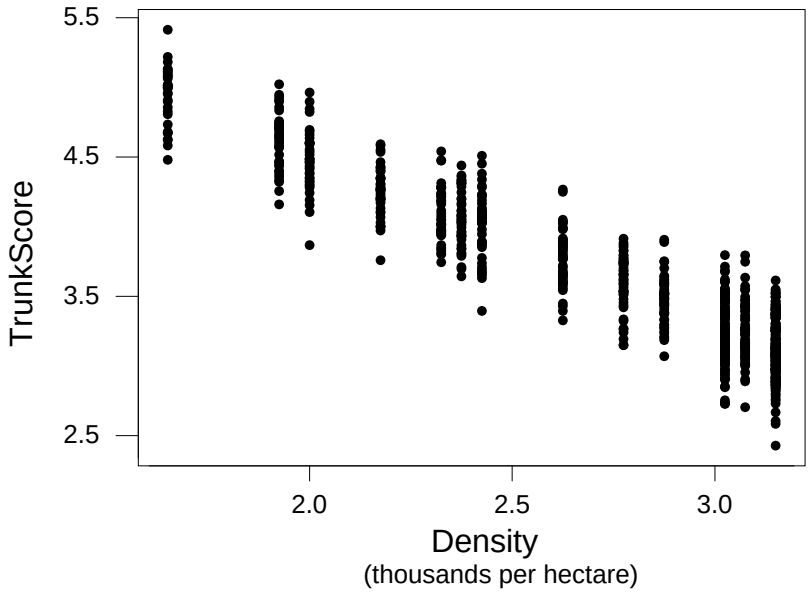

Figure 9: Scatter plot of **TrunkScore** against **Density**

Regression Analysis for TrunkScore on Density

```
The regression equation is
TrunkScore = 6.90 - 1.20 Density

Predictor       Coef      StDev          T          P
Constant     6.89701    0.04279     161.20      0.000
Density     -1.19678    0.01545     -77.48      0.000

S = 0.2055      R-Sq = 87.4%    R-Sq(adj) = 87.3%
```

Table 14: Regression analysis of the relationship between **TrunkScore** and **Density**

Some model checking plots are shown on the next page.

Response is TrunkScore

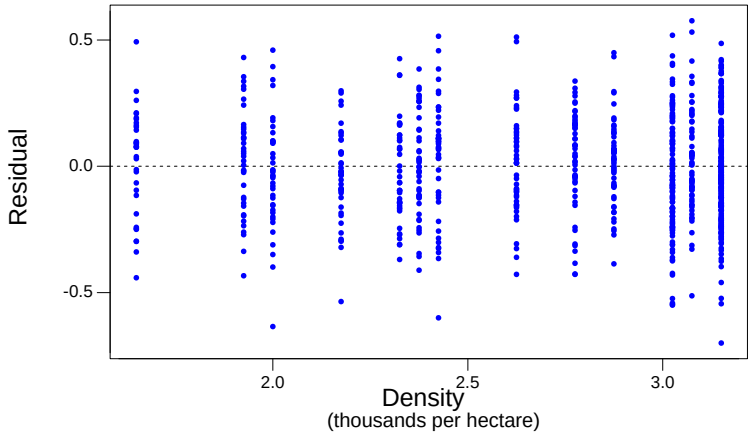

Figure 10: Scatter plot of **Residual** against **Density**

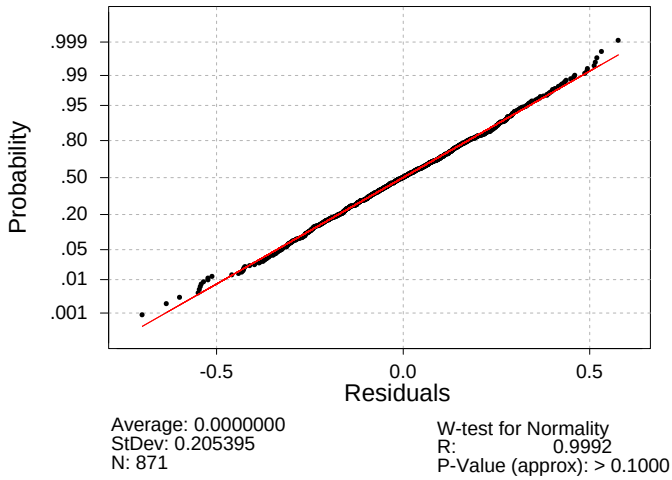

Figure 11: Normality plot of **Residual**

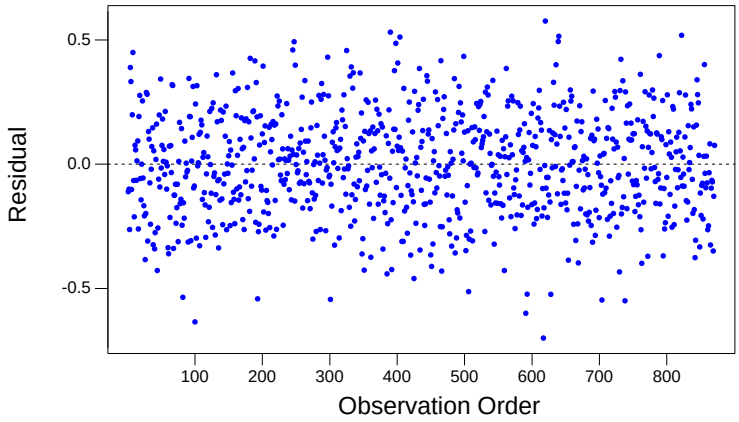

Figure 12: Scatter plot of **Residual** against **Observation Order**

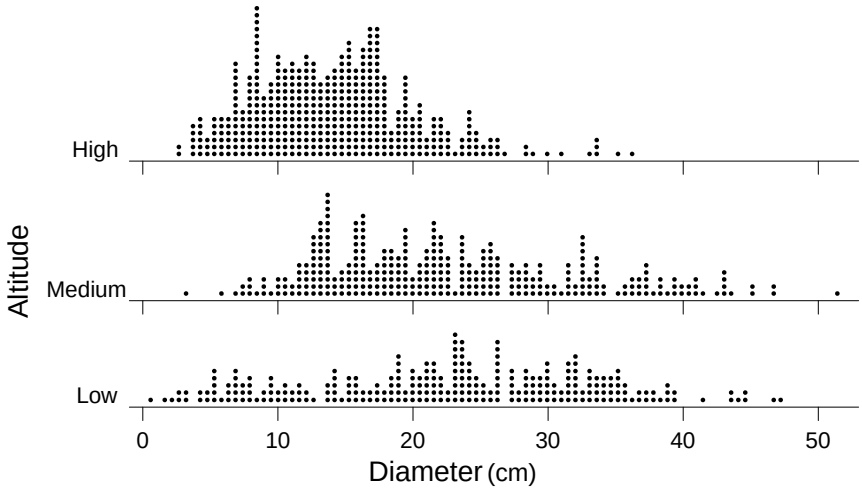

Figure 13: Dot plot of **Diameter** by **Altitude**

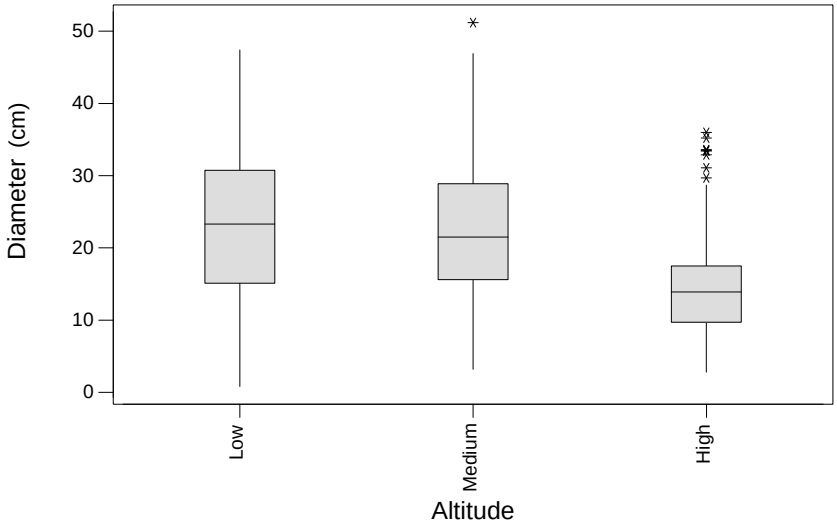

Figure 14: Box plot of **Diameter** by **Altitude**

Descriptive Statistics

Variable	Altitude	N	Mean	Median	StDev
Diameter	Low	210	22.688	23.300	10.408
	Medium	273	22.791	21.500	9.406
	High	388	14.196	13.900	6.005

Table 15: MINITAB output, descriptive statistics of **Diameter** by **Altitude**

One-way Analysis of Variance

```
Analysis of Variance for Diameter
Source     DF        SS        MS        F        P
Altitude  df1   17926.5    8963.3       fo   0.000
Error     df2   70677.5      70.9
Total      df   88604.0

                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level       N      Mean     StDev  ------+---------+---------+---------+-
Low       210    22.688    10.408                              (---*--)
Medium    273    22.791     9.406                               (--*--)
High      388    14.196     6.005   (-*--)
                                  ------+---------+---------+---------+-
Pooled StDev = 8.420                  15.0      18.0      21.00     24.0

Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0195

Critical value = 3.31

Intervals for (column level mean) - (row level mean)

                High           Low

       Low    -10.053
               -6.932

       Medium -10.054         -1.765
               -7.135          1.560
```

Table 16: One-way analysis of variance of **Diameter** by **Altitude**

Appendix E: Fungi Data

Questions 60 to **63** refer to the information given in this appendix.

Hypholoma Capnoides is a pleasant tasting fungus (mushroom) which looks very much like the generally taller, poisonous fungus *Sulphur Tuft*. It has been suggested that the two fungi can be distinguished by use of the following guide: if the fungus is shorter than the threshold height of 8cm then classify it as the edible *Hypholoma Capnoides*, but if the fungus is taller than 8cm then classify it as the poisonous *Sulphur Tuft*.

Let X_H be the height of a *Hypholoma Capnoides* mushroom and X_S be the height of a *Sulphur Tuft* fungus. X_H is well modelled by a Normal distribution with a mean of 6.5cm and a standard deviation of 1.76cm ($X_H \sim$ Normal(μ_H = 6.5cm, σ_H = 1.76cm)) whereas X_S is well modelled by a Normal distribution with mean 9.5cm and a standard deviation 1.25cm ($X_S \sim$ Normal(μ_S = 9.5cm, σ_S = 1.25cm)). Assume X_H and X_S are independent random variables.

Table 17 below shows a selection of probabilities from the distributions of X_H and X_S.

Normal with mean = 6.5 and standard deviation = 1.76		Normal with mean = 9.5 and standard deviation = 1.25	
x	pr(X<=x)	x	Pr(X<=x)
2.4	0.01	6.6	0.01
4.2	0.10	8.0	0.12
5.0	0.20	8.5	0.20
5.3	0.25	8.7	0.25
5.5	0.30	8.8	0.30
6.1	0.40	9.2	0.40
6.5	0.50	9.5	0.50
6.9	0.60	9.8	0.60
7.4	0.70	10.2	0.70
7.7	0.75	10.3	0.75
8.0	0.80	10.6	0.80
8.8	0.90	11.1	0.90
10.6	0.99	12.4	0.99

Table 17: A selection of probabilities from Normal(6.5, 1.76) and Normal(9.5, 1.25) distributions.

ANSWERS:

1. (5)	2. (4)	3. (1)	4. (2)	5. (2)
6. (3)	7. (5)	8. (2)	9. (2)	10. (4)
11. (4)	12. (5)	13. (3)	14. (3)	15. (5)
16. (3)	17. (2)	18. (5)	19. (5)	20. (4)
21. (2)	22. (5)	23. (3)	24. (1)	25. (3)
26. (1)	27. (5)	28. (4)	29. (1)	30. (4)
31. (1)	32. (3)	33. (4)	34. (4)	35. (4)
36. (2)	37. (5)	38. (3)	39. (3)	40. (3)
41. (1)	42. (1)	43. (5)	44. (1)	45. (3)
46. (2)	47. (1)	48. (5)	49. (2)	50. (2)
51. (1)	52. (1)	53. (3)	54. (1)	55. (3)
56. (5)	57. (2)	58. (5)	59. (1)	60. (2)
61. (4)	62. (4)	63. (4)	64. (1)	65. (4)

FORMULAE

Median Position $= \dfrac{n+1}{2}$

Distributions

In general: $\text{sd}(X) = \sqrt{\text{E}(X-\mu_X)^2}$

If X is a **discrete random variable**:

$$\mu_X = \text{E}(X) = \sum x_i \text{pr}(X = x_i) \qquad \text{sd}(X) = \sqrt{\sum (x_i - \mu_X)^2 \text{pr}(X = x_i)}$$

$\boldsymbol{X} \sim$ **Binomial** (n, p) $\quad \text{E}(X) = np \qquad \text{sd}(X) = \sqrt{np(1-p)}$

$\boldsymbol{X} \sim$ **Poisson** (λ) $\quad \text{E}(X) = \lambda \qquad \text{sd}(X) = \sqrt{\lambda}$

$\boldsymbol{X} \sim$ **Normal** (μ, σ) $\quad \text{E}(X) = \mu \qquad \text{sd}(X) = \sigma$

Combining random variables

For any constants a and b:

$$\text{E}(aX + b) = a\text{E}(X) + b \qquad \text{sd}(aX + b) = |a|\text{sd}(X)$$

If X_1 and X_2 are independent random variables:

$$\text{E}(a_1X_1 + a_2X_2) = a_1\text{E}(X_1) + a_2\text{E}(X_2)$$

$$\text{sd}(a_1X_1 + a_2X_2) = \sqrt{a_1^2\text{sd}(X_1)^2 + a_2^2\text{sd}(X_2)^2}$$

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ and standard deviation σ:

$$\text{E}(X_1 + X_2 + \ldots + X_n) = n\mu$$

$$\text{sd}(X_1 + X_2 + \ldots + X_n) = \sqrt{n}\sigma$$

Sampling distributions

$$\text{E}(\overline{X}) = \mu, \quad \text{sd}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

$$\text{E}(\widehat{P}) = p, \quad \text{sd}(\widehat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

Standard error of a difference (independent estimates)

$$\text{se}(\widehat{\theta}_1 - \widehat{\theta}_2) = \sqrt{\text{se}(\widehat{\theta}_1)^2 + \text{se}(\widehat{\theta}_2)^2}$$

Confidence intervals and *t*-tests

Confidence interval: $\quad estimate \pm t\, standard\, errors$

$$\widehat{\theta} \pm t \times \text{se}(\widehat{\theta})$$

t-test statistic: $\quad t_0 = \dfrac{estimate - hypothesised\, value}{standard\, error}$

$$t_0 = \frac{\widehat{\theta} - \theta_0}{\text{se}(\widehat{\theta})}$$

Applications

Mean $\boldsymbol{\mu_X}$: $\quad \widehat{\theta} = \overline{x}, \quad \text{se}(\overline{x}) = \dfrac{s_X}{\sqrt{n}}, \quad df = n - 1$

Proportion $\boldsymbol{p}$: $\quad \widehat{\theta} = \widehat{p}, \quad \text{se}(\widehat{p}) = \sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}, \quad df = \infty$

Difference between two means $\boldsymbol{\mu_1 - \mu_2}$: $\quad \widehat{\theta} = \overline{x}_1 - \overline{x}_2$ (independent samples),

$$\text{se}(\overline{x}_1 - \overline{x}_2) = \sqrt{\text{se}(\overline{x}_1)^2 + \text{se}(\overline{x}_2)^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad df = \text{Min}(n_1 - 1, n_2 - 1)$$

Difference in proportions $\boldsymbol{p_1 - p_2}$: $\quad \widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$ with

(a) ***Proportions from two independent samples*** of sizes n_1, n_2 respectively

$$\text{se}(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \qquad df = \infty$$

(b) ***One sample of size n, several response categories***

$$\text{se}(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{\widehat{p}_1 + \widehat{p}_2 - (\widehat{p}_1 - \widehat{p}_2)^2}{n}} \qquad df = \infty$$

(c) ***One sample of size n, many yes/no items***

$$\text{se}(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{\text{Min}(\widehat{p}_1 + \widehat{p}_2, \widehat{q}_1 + \widehat{q}_2) - (\widehat{p}_1 - \widehat{p}_2)^2}{n}} \qquad df = \infty$$

where $\widehat{q}_1 = 1 - \widehat{p}_1$ and $\widehat{q}_2 = 1 - \widehat{p}_2$

The *F*-test (ANOVA)

$$f_0 = \frac{s_B^2}{s_W^2} \quad df_1 = k - 1 \quad df_2 = n_{\text{tot}} - k$$

The Chi-square test

$$x_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

For one-way tables:

$$df = J - 1$$

For two-way tables:

$$\text{Expected count in cell } (i, j) = \frac{R_i C_j}{n}$$

$$df = (I - 1)(J - 1)$$

Regression

Fitted least-squares regression line: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

Inference about the intercept, β_0, and the slope, β_1: $df = n - 2$
