THE UNIVERSITY OF AUCKLAND

STATS 101
STATS 101G
STATS 108

SECOND SEMESTER, 2014
Campus: CITY

VERSION 1

STATISTICS

Introduction to Statistics
Statistics for Commerce

(Time allowed: THREE hours)

ANSWERS ON PAGE 36

NOTE:

* **This examination consists of 50 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:

* **Appendix A:** *ACTN3* **Genotype Data** pages 27 and 28
* **Appendix B: Inkjet Printer Data** pages 29 and 30
* **Appendix C: Fish Ventilation Data** pages 31 and 32
* **Appendix D: Pharmacy Data** pages 33 and 34
* **Appendix E: NFL Data** pages 35 and 36

BLANK PAGE

Questions 1 to **12** refer to the information in **Appendix A**, pages 27 and 28.

Use Table 1, page 27, to answer **Questions 1** and **2**.

1. The proportion of the non-athletes who possess the RX genotype is approximately:

(1) 0.52
(2) 0.59
(3) 0.62
(4) 0.31
(5) 0.49

2. Based on the figures in Table 1, page 27, how many times as likely is it that a sprinter has the RR genotype compared to an endurance athlete?

(1) About 0.9 times as likely
(2) About 1.8 times as likely
(3) About 1.1 times as likely
(4) About 0.6 times as likely
(5) About 1.6 times as likely

CONTINUED

Questions 3 and **4** refer to Table 1, page 27, and the following additional information.

Let:

p_{E} be the underlying proportion of elite endurance athletes, in Australia, who possess the RX genotype

and

p_{S} be the underlying proportion of elite sprinters, in Australia, who possess the RX genotype.

3. The sampling situation associated with $\text{se}(\widehat{p}_{\text{E}} - \widehat{p}_{\text{S}})$ can be described as:

(1) one single sample of size 737, several response categories.
(2) two independent samples, one of size 194 and one of size 107.
(3) one single sample of size 737, many yes/no items.
(4) one single sample of size 362, several response categories.
(5) two independent samples, one of size 88 and one of size 48.

4. A 95% confidence interval for $p_{\text{E}} - p_{\text{S}}$ is $(-0.1124, 0.1224)$.

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

With 95% confidence we estimate that, in Australia, the underlying proportion of elite endurance athletes who possess the RX genotype is:

(1) 0.005 greater than the corresponding proportion of elite sprinters, with a margin of error of 0.235.
(2) 0.11 and the corresponding proportion of elite sprinters is 0.12.
(3) somewhere between 0.11 less than and 0.12 greater than the corresponding proportion of elite sprinters.
(4) somewhere between 0.11 and 0.12 less than the corresponding proportion of elite sprinters.
(5) 0.235 greater than the corresponding proportion of elite sprinters, with a margin of error of 0.117.

CONTINUED

Questions 5 to **12** refer to the information on page 28.

5. Which **one** of the following is an **incorrect** null hypothesis for the Chi-square test described on page 28?

   (1) The underlying distribution of **Genotype** is the same for each level of **Group**.

   (2) There is no association between **Genotype** and **Group**.

   (3) In Australia, athletic type is not related to *ACTN3* genotype.

   (4) In Australia, athletic type is independent of *ACTN3* genotype.

   (5) The underlying distribution of **Group** is the same for each level of **Genotype**.

6. Which **one** of the following statements gives the **correct** justification concerning the validity of this Chi-square test?

   (1) There is no cause for concern because no cells have an expected count less than 5.

   (2) There is concern because at least one cell has a cell contribution to the Chi-square test statistic that is less than 1.

   (3) There is no cause for concern because no cells have an observed count less than 5.

   (4) There is concern because at least five cells have a cell contribution to the Chi-square test statistic that is less than 5.

   (5) There is concern because one cell has an observed count less than 10.

Questions 7 to **12** assume that the Chi-square test is valid.
(Note this may not be true.)

7. Consider the cell in Table 2, page 28, for non-athletes who possess the RX genotype. Under the null hypothesis, the expected count for this cell is:

   (1) 81.89

   (2) 226.00

   (3) 120.67

   (4) 214.15

   (5) 145.33

8. Consider the cell in Table 2, page 28, for endurance athletes who possess the XX genotype. This cell's contribution to the Chi-square test statistic, to 2 decimal places, is:

   (1) 0.32

   (2) 0.11

   (3) 11.27

   (4) 3.65

   (5) 2.76

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

   For sprinters, under the null hypothesis, we would expect the proportions to be about:

   (1) $\frac{1}{3}$ for all three levels of **Genotype**.

   (2) $\frac{53}{107}$, $\frac{48}{107}$ and $\frac{6}{107}$ for the respective levels of **Genotype**.

   (3) $\frac{53}{243}$, $\frac{48}{362}$ and $\frac{6}{132}$ for the respective levels of **Genotype**.

   (4) $\frac{243}{737}$, $\frac{362}{737}$ and $\frac{132}{737}$ for the respective levels of **Genotype**.

   (5) $\frac{107}{737}$ for all three levels of **Genotype**.

10. Based on the results of this Chi-square test, which **one** of the following statements is **true**?

    (1) The probability that the alternative hypothesis is false is approximately 0.000.

    (2) The probability that the null hypothesis is true is approximately 0.000.

    (3) If the alternative hypothesis is true, then the probability of getting a test statistic at least as large as 24.805 is approximately 0.000.

    (4) If the null hypothesis is true, then the probability of getting a test statistic of 24.805 or smaller is approximately 0.000.

    (5) If the null hypothesis is true, then the probability of getting a test statistic at least as large as 24.805 is approximately 0.000.

11. Which **one** of the following statements is **not** a valid conclusion of this Chi-square test?

There is very strong evidence:

(1) that the underlying distribution of **Genotype** is different for each level of **Group**.

(2) that there is a relationship between the variables **Group** and **Genotype**.

(3) that the underlying distribution of **Genotype** is not the same for each level of **Group**.

(4) that the variables **Group** and **Genotype** are not independent.

(5) of a link between the variables **Group** and **Genotype**.

12. Which **one** of the following statements gives the **best** reason for the small *P-value* of 0.000?

(1) The number of non-athletes possessing the XX genotype was very close to the number expected under the null hypothesis.

(2) The number of sprinters possessing the RR genotype was far more than expected and the number of sprinters possessing the XX genotype was far fewer than expected under the null hypothesis.

(3) The number of non-athletes possessing the RX genotype was far greater than the number of sprinters or endurance athletes possessing the RX genotype.

(4) The number of non-athletes was far greater than the number of sprinters or the number of endurance athletes.

(5) The number of endurance athletes possessing the XX genotype was far more than expected under the null hypothesis.

Questions 13 to **23** refer to the information in **Appendix B**, pages 29 and 30.

13. Based on the scatter plot in Figure 4, page 29, which **one** of the following statements is **false**?

(1) An increase in printing rate tends to be associated with an increase in price.

(2) The plot indicates a linear association between **Price** and **PPM**.

(3) The inkjet printer with a printing rate of 4.1 ppm and a price of $199 is considered an outlier.

(4) The relationship between **Price** and **PPM** is moderate.

(5) Of the 20 inkjet printers, the one with the slowest printing rate is the cheapest.

14. Based on the residual plot in Figure 5, page 30, which **one** of the following statements is **false**?

(1) The plot indicates that we should have no concerns about the assumption of linearity.

(2) The plot does not provide information about the assumption of independence of the errors.

(3) The plot indicates that we should have no concerns about the assumptions of Normality of the errors and constant spread of the errors.

(4) The plot does not give us any concern about using the linear model.

(5) The plot indicates non-constant spread and so we should be concerned about using the linear model.

Questions 15 to **22** assume that the simple linear regression analysis on **Price** and **PPM** is valid.
(Note this may not be true.)

15. The equation of the least squares regression line for this analysis is:

- **(1)** Predicted **Price** $= 90.878 - 94.222 \times$ **PPM**
- **(2)** Predicted **Price** $= -94.222 + 90.878 \times$ **PPM**
- **(3)** Predicted **PPM** $= -94.222 + 0.740 \times$ **Price**
- **(4)** Predicted **PPM** $= -94.222 + 90.878 \times$ **Price**
- **(5)** Predicted **Price** $= -94.222 + 0.740 \times$ **PPM**

16. The residual for the inkjet printer that prints 4.1 ppm and has a price of \$199 is:

- **(1)** $-\$79.38$
- **(2)** \$173.60
- **(3)** \$79.38
- **(4)** $-\$173.60$
- **(5)** \$278.38

17. For two inkjet printers like those in the study that have a printing rate that differs by 0.5 ppm, this regression analysis predicts that their prices would differ by:

- **(1)** \$22
- **(2)** \$47
- **(3)** \$45
- **(4)** \$2632
- **(5)** \$91

CONTINUED

18. In a two-tailed t-test for no linear relationship between **Price** and **PPM** the correct hypotheses are:

- **(1)** H_0: $\widehat{\beta}_1 = 0$ H_1: $\widehat{\beta}_1 \neq 0$
- **(2)** H_0: $\beta_1 = 0$ H_1: $\beta_1 \neq 0$
- **(3)** H_0: $\beta_1 \neq 0$ H_1: $\beta_1 = 0$
- **(4)** H_0: $\beta_0 = 0$ H_1: $\beta_0 \neq 0$
- **(5)** H_0: $\widehat{\beta}_0 \neq 0$ H_1: $\widehat{\beta}_0 = 0$

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

- **(1)** There is very strong evidence against a linear relationship between **Price** and **PPM**.
- **(2)** There is very strong evidence of a linear relationship between **Price** and **PPM**.
- **(3)** There is evidence of a very strong non-linear relationship between **Price** and **PPM**.
- **(4)** There is evidence of a very strong linear relationship between **Price** and **PPM**.
- **(5)** There is very strong evidence of a non-linear relationship between **Price** and **PPM**.

20. For printers like those in the study, which **one** of the following statements is **true**?

- **(1)** With 95% confidence we estimate that, on average, a \$1 increase in the price of a printer is associated with an increase in the printing rate of somewhere between 50 and 132 ppm.
- **(2)** With 95% confidence we estimate that, on average, a one page per minute increase in the printing rate is associated with an increase in the price of a printer of \$91, with a margin of error of approximately \$82.
- **(3)** On average, a \$1 increase in the price of a printer is associated with an increase in the printing rate of 91 ppm.
- **(4)** With 95% confidence we estimate that, on average, a one page per minute decrease in the printing rate is associated with an increase in the price of a printer of somewhere between \$50 and \$132.
- **(5)** It is plausible that, on average, a one page per minute increase in the printing rate is associated with an increase in the price of a printer of \$64.

CONTINUED

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

With 95% confidence, it is estimated that for printers like those in the study that have a printing rate of 2.9 ppm, the mean price of printers is somewhere between:

(1) \$43 and \$295
(2) \$43 and \$197
(3) \$142 and \$197
(4) \$142 and \$295
(5) \$129 and \$199

22. Suppose it was found that there was a mistake in the recorded data for the printer that had a printing rate of 4.1 ppm and a price of \$199. The printing rate was correct but the price should have been \$299. The analysis was re-run with the correct values, giving new estimates for the slope of the regression line and the correlation coefficient.

Which **one** of the following statements about the new estimates relative to the original estimates is **true**?

(1) The new slope estimate will be slightly higher and the new sample correlation coefficient will be slightly further from 1.
(2) The new slope estimate will be slightly lower and the new sample correlation coefficient will be closer to 1.
(3) The new slope estimate will be slightly lower and the new sample correlation coefficient will be slightly further from 1.
(4) The new slope estimate will be the same and the new sample correlation coefficient will be closer to 1.
(5) The new slope estimate will be slightly higher and the new sample correlation coefficient will be closer to 1.

23. Which **one** of the following forms of analysis could be used to investigate the relationship between **CostBW** and **CostColour** as defined on page 29?

(1) Simple linear regression
(2) A Chi-square test for independence
(3) A t-test for no difference between two means
(4) A Chi-square test for goodness of fit
(5) A one-way analysis of variance F-test

CONTINUED

Questions 24 to **27** refer to the following information.

Researchers in the United States studied whether children who were breastfed differed from those who were not breastfed. The subjects were 322 children whose parents volunteered them to be part of the study. 237 of these children had been breastfed during infancy and 85 had not. At age 4 years each child was measured for a General Cognitive Index (GCI) score which is an overall measure of a child's level of intellectual functioning. Higher GCI scores indicate higher intellectual functioning. The GCI scores and whether or not each child had been breastfed during infancy were recorded. Plots of the data and the difference between the group means are shown in Figure 1. Data sourced from Tintle *et al.* (2014).

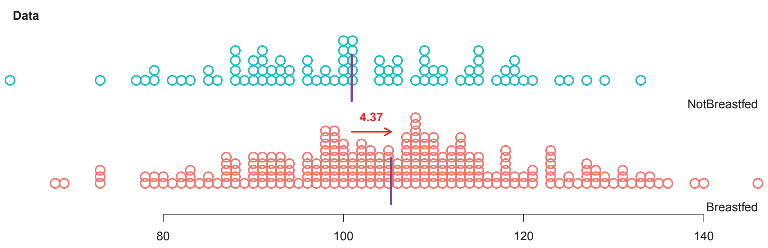

Figure 1: GCI scores by group

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

(1) Blocking existed in this study; the children were blocked on whether they were breastfed during infancy or not.
(2) Blinding existed in this study as the children, at age 4 years, would not have known whether they were breastfed during infancy or not.
(3) For this study to be considered an experiment, the children should have been randomly allocated to the two groups according to their GCI scores.
(4) For this study, the groups were determined by which children who, during infancy, happened to be breastfed or not.
(5) The placebo effect would have to be taken into consideration when GCI scores of the children were measured.

CONTINUED

25. If the observed difference between the two groups had happened by chance alone, which **one** of the following statements is **false**?

(1) The mean GCI score would have been the same for both groups.

(2) The child whose GCI score was lowest would still have had the lowest GCI score regardless of whether they had been breastfed during infancy or not.

(3) The observed difference between the group means can be explained as the result of which children just happened to be in each group.

(4) Differences between the two groups can be simulated by randomly re-allocating the observed GCI scores to the two groups.

(5) There would have been no reason to expect lower GCI scores in one group and larger GCI scores in the other group.

Questions 26 and **27** refer to the following additional information.

A one-tailed randomisation test was performed on the data. The test output in Figure 2 shows a re-randomisation distribution for the difference between the mean GCI score for each group and the observed difference between the two group means (4.37) as displayed in Figure 1, page 12.

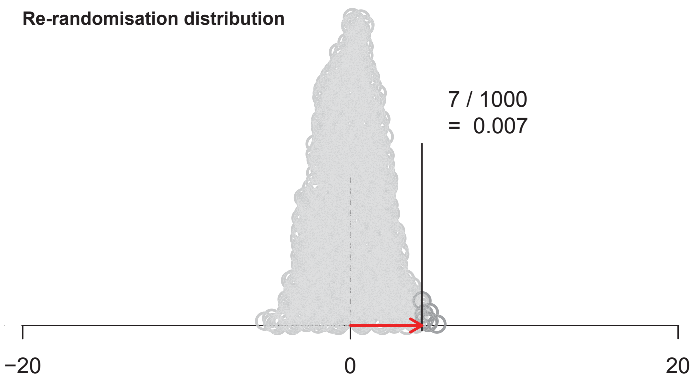

Figure 2: Re-randomisation distribution for the breastfeeding data

26. Which **one** of the following statements about this test output is **true**?

(1) If the study was repeated 1000 times, then about 7 of those 1000 studies would produce an observed difference between the group means of at least 4.37.

(2) At least 7 out of 1000 re-randomisations under chance alone produced a difference between the group means of 4.37.

(3) 7 out of 1000 re-randomisations under chance alone produced a difference between the group means of at least 4.37.

(4) If 1000 different researchers repeated this study, then about 7 of those 1000 studies would produce an observed difference between the group means of at least 4.37.

(5) At least 7% of the re-randomisations under chance alone produced a difference between the group means of at least 4.37.

27. Which **one** of the following statements is a **not** a valid conclusion from the randomisation test result?

(1) We may claim that breastfeeding is the reason children breastfed during infancy will have, on average, higher GCI scores at age 4 years than children not breastfed during infancy.

(2) For children in the study, we may not claim that the connection between breastfeeding during infancy and GCI scores at age 4 years is causal.

(3) We cannot claim that the observed difference between the two group means is due to the fact that one group of children were breastfed during infancy whilst the other group were not.

(4) We cannot claim that there is a link between breastfeeding during infancy and GCI scores for all children at age 4 years.

(5) We may claim that for the children in the study there is a link between breastfeeding during infancy and GCI scores at age 4 years.

Questions 28 to **33** refer to the information in **Appendix C**, pages 31 and 32.

28. Which **one** of the following statements about the data is **false**?

(1) The distribution of gill beat rate for the medium calcium group is similar to that for the high calcium group.

(2) The range for all 90 fish is 65 gill beats per minute.

(3) The low calcium group has the largest range of gill beat rates.

(4) The gill beat rates for the low calcium group are left skewed.

(5) The gill beat rate is highest, on average, for the low calcium group.

29. Correct hypotheses for the F-test described on page 32 are:

(1) H_0: The underlying mean gill beat rate is the same for all 3 calcium level groups.
H_1: The underlying mean gill beat rate is different for all 3 calcium level groups.

(2) H_0: The group mean gill beat rate is the same for all 3 calcium level groups.
H_1: The group mean gill beat rate is different for at least one of the 3 calcium level groups.

(3) H_0: The underlying mean gill beat rate is the same for all 3 calcium level groups.
H_1: The group mean gill beat rate is different for at least one of the 3 calcium level groups.

(4) H_0: The underlying mean gill beat rate is the same for all 3 calcium level groups.
H_1: The group mean gill beat rate is different for all 3 calcium level groups.

(5) H_0: The underlying mean gill beat rate is the same for all 3 calcium level groups.
H_1: The underlying mean gill beat rate is different for at least one of the 3 calcium level groups.

In **Questions 30** to **33** the use of the F-test is appropriate.

30. Which **one** of the following statements is **not** a justification concerning the assumptions of this F-test?

(1) There is no concern about the Normality assumption because the group size is large enough for the test to withstand any non-Normal features in the data.

(2) There is no concern with the assumption that the means of the underlying distributions are equal since the ratio of the largest group mean to the smallest is less than 2.

(3) Since the fish were randomly allocated to the 3 calcium groups there is no concern about the independence of the observations within each of the 3 groups.

(4) Since the fish were randomly allocated to the 3 calcium groups there is no concern about the independence of the 3 groups.

(5) There is no concern with the assumption that the standard deviations of the underlying distributions are equal since the ratio of the largest group standard deviation to the smallest is less than 2.

31. Using the output from Table 5, page 32, which **one** of the following statements is **true**?

(1) $\text{pr}(F \geq 4.648) = 0.006$ where $F \sim \text{F}(df_1 = 2, df_2 = 87)$

(2) $\text{pr}(F \geq 0.006) = 4.648$ where $F \sim \text{F}(df_1 = 2, df_2 = 87)$

(3) $\text{pr}(F \geq 0.012) = 4.648$ where $F \sim \text{F}(df_1 = 2, df_2 = 87)$

(4) $\text{pr}(F \leq 4.648) = 0.012$ where $F \sim \text{F}(df_1 = 2, df_2 = 87)$

(5) $\text{pr}(F \geq 4.648) = 0.012$ where $F \sim \text{F}(df_1 = 2, df_2 = 87)$

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

(1) There is not enough variability within the three groups relative to the variability between the three group means to make the alternative hypothesis plausible.

(2) There is not enough variability between the three group means relative to the variability within the three groups to make the null hypothesis plausible.

(3) There is too much variability within the three groups relative to the variability between the three group means to make the null hypothesis plausible.

(4) There is too much variability between the three group means relative to the variability within the three groups to make the null hypothesis plausible.

(5) There is not enough variability between the three group means relative to the variability within the three groups to reject the null hypothesis.

33. Which **one** of the following conclusions made from the Tukey multiple comparisons output in Table 5, page 32, is **true**?

(1) The underlying mean gill beat rate is the lowest for fish in the high calcium group.

(2) The underlying mean gill beat rate is the highest for fish in the low calcium group.

(3) We can ascertain that there is a significant difference (at the 5% level) between the mean gill beat rate for the fish in the low calcium group and that for the fish in the other two groups, but not which group has the highest underlying mean gill beat rate.

(4) It cannot be ascertained which calcium group has the highest, nor the lowest, underlying mean gill beat rate.

(5) The underlying mean gill beat rate is the highest for fish in the low calcium group followed by the medium calcium group, with fish in the high calcium group having the lowest.

Questions 34 to **43** refer to the information in **Appendix D**, pages 33 and 34.

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

(1) We would expect another sample of pharmacists (of the same size from the same population) to have a difference between sample means of 2.78.

(2) We would expect another bootstrap distribution from the same sample to be similar to the one obtained in Figure 8, page 33.

(3) In the 1000 re-samples most of the mean compassion satisfaction scores for the females were greater than those for the males.

(4) While the plots from a new sample of pharmacists (of the same size from the same population) may look quite different to those shown in the Sample section of Figure 8, page 33, it would be a fairly safe bet that its resulting bootstrap confidence interval would contain the difference between the population means.

(5) The mean compassion satisfaction score for the female sample is 2.78 higher than that for the male sample.

35. Which **one** of the following is a **correct** interpretation of the bootstrap confidence interval shown in Figure 8, page 33.

For pharmacists who work in a city, it is a fairly safe bet that:

(1) the sample mean compassion satisfaction score for both males and females is somewhere between -0.1 and 5.7.

(2) the population mean compassion satisfaction score for both males and females is somewhere between -0.1 and 5.7.

(3) the sample mean compassion satisfaction score for males could be the same as that for females.

(4) the population mean compassion satisfaction score for males is somewhere between 5.7 lower than and 0.1 higher than that for females.

(5) the population mean compassion satisfaction score for females is somewhere between 5.7 lower than and 0.1 higher than that for males.

Questions 36 to **40** refer to the information on page 34.

36. Consider the summary statistics in Table 6, page 34, for the sample of the 142 **rural** pharmacists and let μ_R be the mean compassion satisfaction score of the population of rural pharmacists.

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

(1) $s_R = 7.537$

(2) 0.632 estimates the variability in sample means of repeated samples of size 142 taken from the population.

(3) Roughly, the average difference between the observations in the sample and the sample mean, $\overline{x}_R$, is 0.632.

(4) An estimate for μ_R is given by 30.99.

(5) $\overline{x}_R = 30.99$

37. In considering the appropriateness of using a two independent sample t-test, which **one** of the followings statements is **true**?

(1) We have concerns about using a t-test because of the unusually large number of outliers in these data.

(2) Without a plot of the differences we are unable to say whether we should have concerns about using a t-test.

(3) We have concerns about using a t-test because the two samples may not be independent.

(4) We have no concerns about using a t-test with respect to the skewness suggested by these data because the total sample size is large enough for the test to withstand the level of skewness shown.

(5) We have no concerns about using a t-test because the standard error of the estimate is sufficiently small.

Questions 38 to **40** assume that the use of a t-test is appropriate.
(Note this may not be true.)

38. In this two independent sample t-test, what is the value of the test statistic?

(1) 0.04

(2) 0.09

(3) 1.99

(4) 0

(5) 1.95

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

(1) There is some evidence that the mean compassion satisfaction score is higher for city pharmacists than for rural pharmacists.

(2) There is some evidence that city pharmacists have higher compassion satisfaction scores than rural pharmacists.

(3) There is some evidence against the mean compassion satisfaction score for city pharmacists being the same as that for rural pharmacists.

(4) There is some evidence that, on average, there is a difference in compassion satisfaction scores between city and rural pharmacists.

(5) The observed difference of 1.915 between the means of the compassion satisfaction scores for the city and rural pharmacists is large enough to be statistically significant at the 5% level.

40. Suppose the 95% confidence interval had been calculated manually (rather than using SPSS). Which **one** of the following statements is **true**?

The 95% confidence interval for the manual calculation will be slightly:

(1) wider than that in Table 6 because the *df* used will be 250.

(2) wider than that in Table 6 because the *df* used will be 108.

(3) wider than that in Table 6 because the *df* used will be 141.

(4) narrower than that in Table 6 because the *df* used will be 250.

(5) narrower than that in Table 6 because the *df* used will be 108.

41. Figure 3 is a histogram of the compassion satisfaction scores for the female pharmacists in the sample.

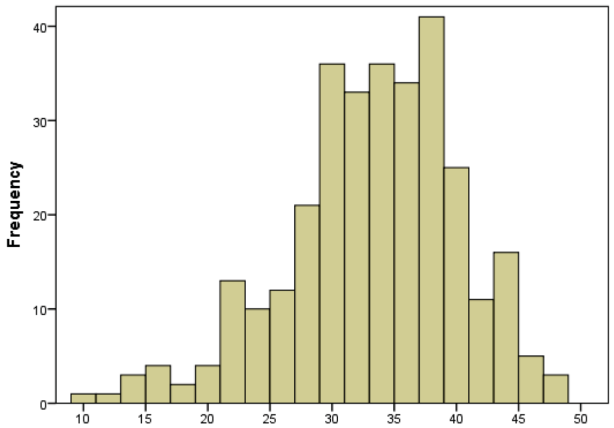

Figure 3: Compassion satisfaction score

Which **one** of the following statements about this histogram is **false**?

(1) The interquartile range is roughly 25.

(2) The range is roughly 40.

(3) The mean is somewhere between 30 and 35.

(4) The distribution is skewed to the lower values.

(5) The standard deviation is less than 12.

42. **No longer examined**

43. Another measure concerning the professional quality of life recorded for the sample of pharmacists was called a **Burnout** score. This was a measure (from 0 to 50) of the feelings of helplessness and difficulties in dealing with work or in doing their job effectively. Suppose we wish to investigate the relationship between **Burnout** and **Compassion satisfaction**.

Which **one** of the following is the **best** tool to explore such a relationship?

(1) Calculating the mean and standard deviation of the scores for both **Burnout** and **Compassion satisfaction** and comparing them.

(2) Side-by-side box plots of **Burnout** and **Compassion satisfaction**, on the same scale.

(3) A scatter plot of **Burnout** against **Compassion satisfaction**.

(4) A two-way table of counts, cross-classified by **Burnout** and **Compassion satisfaction**.

(5) Bar graphs of proportions showing the differences in the distributions of **Burnout** for each level of **Compassion satisfaction**.

44. In reference to polls and surveys which **one** of the following statements is **false**?

(1) Using random samples in statistical surveys is an attempt to obtain a representative sample.

(2) Increasing the sample size will not reduce the effects of nonsampling errors.

(3) Under simple random sampling, each unit in the population has an equal chance of being chosen.

(4) Sampling errors have the potential to be smaller in larger samples than in smaller samples.

(5) When taking a sample, the aim is to have low bias and low precision.

45. Which **one** of the following statements about sample-to-population (statistical) inference is **false**?

(1) Quoting a margin of error is a way of communicating the uncertainty due to sampling variability.

(2) A parameter is a numerical characteristic of a population or distribution.

(3) A confidence interval is used to give a range of plausible values for the sample estimate.

(4) The process of using sample data to make useful statements about a population is called sample-to-population (statistical) inference.

(5) An estimate is a known quantity calculated from the sample data to estimate an unknown parameter.

Questions 46 to **49** refer to the information in **Appendix E**, pages 35 and 36.

46. The correct hypotheses for this t-test are:

(1) H_0: $\tilde{\mu}_{\text{Diff}} = 0$ H_1: $\tilde{\mu}_{\text{Diff}} \neq 0$

(2) H_0: $\mu_{\text{Diff}} = 0$ H_1: $\mu_{\text{Diff}} > 0$

(3) H_0: $\mu_{\text{Diff}} = 0$ H_1: $\mu_{\text{Diff}} \neq 0$

(4) H_0: $\mu_{\text{Diff}} = 0$ H_1: $\mu_{\text{Diff}} < 0$

(5) H_0: $\tilde{\mu}_{\text{Diff}} = 0$ H_1: $\tilde{\mu}_{\text{Diff}} > 0$

47. Which **one** of the following statements about this t-test is **true**?

(1) The outliers in the box plot in Figure 11, page 35, give us concern about using the t-test.

(2) To help us decide on the validity of the test with respect to the Normality assumption we need to look at the box plots in both Figures 10 and 11, page 35.

(3) To help us decide on the validity of the test with respect to the assumption of independence we need only look at the box plots in Figure 10, page 35.

(4) We can think of this test as a one sample t-test on the differences obtained from subtracting each game's away score from its corresponding home score.

(5) **This option is no longer examined.**

Questions 48 and **49** assume that the use of a t-test is appropriate.
(Note this may not be true.)

48. Which **one** of the following statements about the *P-value* is **true**?

If there really is no difference between home and away scores on average, then the probability of:

(1) getting a sample mean difference (**Home** − **Away**) of 3.266 is 0.001.

(2) getting a sample mean difference (**Home** − **Away**) of at least 3.266 is 0.001.

(3) getting a sample mean difference (**Home** − **Away**) of at least 3.266 is 0.0005.

(4) the underlying mean difference (**Home** − **Away**) being 0 is 0.0005.

(5) the underlying mean difference (**Home** − **Away**) being 0 is 0.001.

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

(1) We have evidence that there is a very large home game advantage, on average.

(2) At the 5% level of significance we can claim that a home game advantage effect exists.

(3) We have very strong evidence of a home game advantage effect.

(4) At the 1% level of significance we can claim that a home game advantage effect exists.

(5) The test result is highly significant.

50. **No longer examined**

BLANK PAGE

INCLUSIONS:

- **Appendix A:** *ACTN3* **Genotype Data** for use in Questions **1** to **12**
- **Appendix B: Inkjet Printer Data** for use in Questions **13** to **23**
- **Appendix C: Fish Ventilation Data** for use in Questions **28** to **33**
- **Appendix D: Pharmacy Data** for use in Questions **34** to **43**
- **Appendix E: NFL Data** for use in Questions **46** to **49**

References

Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., and Lock, D. F. (2013). *Statistics Unlocking the Power of Data*. Wiley.

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., and Vanderstop, J. (2014). *Statistical Investigtions*. Wiley.

Yang, N., MacArthur, D. G., Gulbin, J. P., Hahn, A. G., Beggs, A. H., Easteal, S., and North, K. (2003). ACTN3 genotype is associated with human elite athletic performance. *American Journal of Human Genetics*, **73**, 627–631.

Appendix A: *ACTN3* Genotype Data

Questions 1 to **12** refer to the information in this appendix.

ACTN3 is a gene responsible for making a protein called α-actinin-3 which functions, almost exclusively, in the fast-twitch muscle fibres of the human body. People can be classified according to which genotype of this gene they possess: RR, RX or XX.

Yang *et al.* (2003) conducted a study in Australia which investigated possible associations between genetics and athletic type. The study was made up of three samples of elite sprinters, elite endurance athletes and a control group of non-athletes, respectively, with each participant being categorised by the *ACTN3* genotype they possessed.

Variables are defined as:

Genotype The *ACTN3* genotype the participant possessed
– RR
– RX
– XX

Group The group (athletic type) the participant belonged to
– Sprint
– Endurance
– Control (non-athletes)

The results are shown in Table 1.

	Genotype			
Group	RR	RX	XX	**Total**
Sprint	53	48	6	107
Endurance	60	88	46	194
Control	130	226	80	436
Total	243	362	132	737

Table 1: Group and genotype

Assume that these 107 elite sprinters, 194 elite endurance athletes and 436 non-athletes are a random sample of all elite sprinters, all elite endurance athletes and all non-athletes in Australia, respectively.

CONTINUED

Questions 5 to **12** refer to the following additional information.

A Chi-square test was conducted to see whether, in Australia, there is an association between genotype and athletic type. The results of this Chi-square test are shown in Table 2.

			Genotype			
			RR	RX	XX	Total
Group	Sprint	Count	53	48	6	107
		Expected count	35.280	52.556	19.164	107.0
		Cell contribution	8.900	0.395	9.043	
	Endurance	Count	60	88	46	194
		Expected count	++	++	34.746	194.0
		Cell contribution	0.246	++	++	
	Control	Count	130	226	80	436
		Expected count	++	++	78.090	436.0
		Cell contribution	1.316	0.655	0.047	
Total		Count	243	362	132	737
		Expected count	243.0	362.0	132.0	737.0

Chi-Square Tests

	Value	df	Sig.
Pearson Chi-Square	24.805	4	.000
Likelihood Ratio	26.678	4	.000
Linear-by-Linear Assocation	.001	1	.978
N of Valid Cases	737		

Note: Some values have been replaced by ++.

Table 2: Chi-square test output

CONTINUED

Appendix B: Inkjet Printer Data

Questions 13 to **23** refer to the information in this appendix.

A small company was interested in replacing their inkjet printers. It was decided to investigate different models of inkjet printer before determining which model they would purchase as a replacement for their current model. The company collected reviews for a sample of 20 multifunction inkjet printers. Information was collected on a large number of variables. Assume that the 20 inkjet printers are a random sample from some underlying population of inkjet printers.
Data sourced from Lock *et al.* (2013).

Four of the variables recorded are defined as:

PPM	The printing rate for a standard set of jobs in pages per minute (ppm)
Price	The retail price of the inkjet printer in US dollars
CostBW	The average cost per page for black and white printing in US cents
CostColour	The average cost per page for colour printing in US cents

A simple linear regression analysis was carried out to investigate whether the printing rate could be used to predict the retail price of an inkjet printer. The results of this analysis and associated plots are shown in Figure 4 below and in Figure 5 and Tables 3 and 4, page 30.

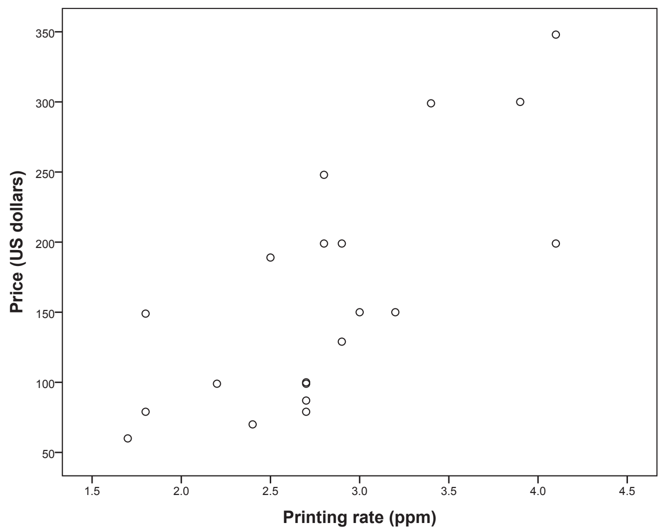

Figure 4: Retail price against printing rate

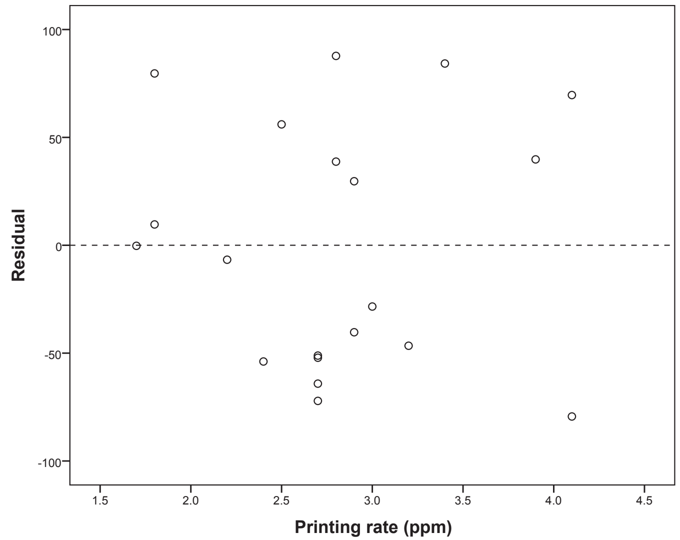

Figure 5: Residual plot

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-94.222	56.398		-1.671	.112	-212.710	24.266
	PPM	90.878	19.488	.740	4.663	.000	49.936	131.820

a. Dependent Variable: Price

Table 3: Simple linear regression output

PPM	Price	LMCI_1	UMCI_1	LICI_1	UICI_1
2.9	199	141.602	197.048	43.239	295.410
2.9	129	141.602	197.048	43.239	295.410

Table 4: Simple linear regression prediction output

Appendix C: Fish Ventilation Data

Questions 28 to **33** refer to the information in this appendix.

Most fish use gills for respiration in water and researchers can observe how fast a fish's gill cover beats to study ventilation (much like we might observe breathing rate for a person). In an experiment to see how water chemistry might affect gill beat rates, 90 fish were randomly assigned to 3 tanks, each with a different level of calcium; Low, Medium and High. One at a time, the fish were put in their allocated tank and after they acclimatised their gill beat rates were recorded. Plots are shown in Figures 6 and 7.
Data sourced from Lock *et al.* (2013).

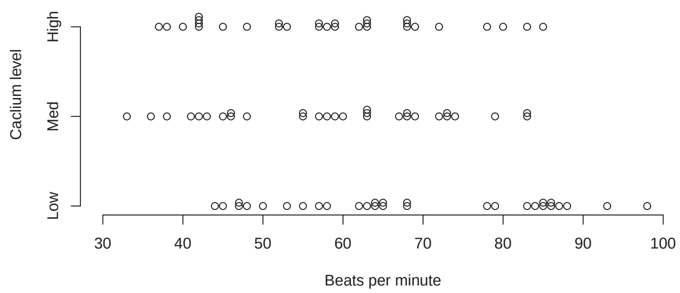

Figure 6: Gill beat rate

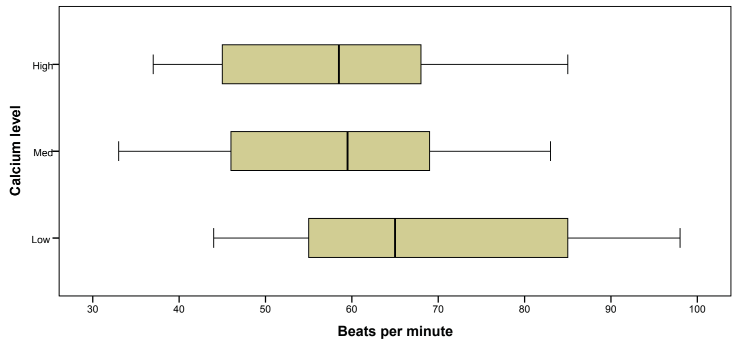

Figure 7: Gill beat rate

CONTINUED

Questions 29 to **33** refer to the following additional information.

The underlying mean for the high calcium group is considered to be the mean gill beat rate if all 90 fish had been put (one at a time) in the high calcium level tank. The other two underlying means are defined similarly.

A one-way analysis of variance (ANOVA) F-test was carried out to see whether there is a relationship between the level of calcium and gill beat rates. The test output is shown in Table 5.

Descriptives

GillRate

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Low	30	68.5000	16.23481	2.96406	62.4378	74.5622	44.00	98.00
Medium	30	58.6667	14.28366	2.60783	53.3331	64.0003	33.00	83.00
High	30	58.1667	13.77675	2.51528	53.0223	63.3110	37.00	85.00
Total	90	61.7778	15.39793	1.62308	58.5527	65.0028	33.00	98.00

ANOVA

GillRate

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2037.222	2	1018.611	4.648	.012
Within Groups	19064.333	87	219.130		
Total	21101.556	89			

Multiple Comparisons

Dependent Variable: GillRate

Tukey HSD

(I) Calcium	(J) Calcium	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Low	Medium	9.83333	3.82213	.031	.7195	18.9471
	High	10.33333	3.82213	.022	1.2195	19.4471
Medium	Low	-9.83333	3.82213	.031	-18.9471	-.7195
	High	.50000	3.82213	.991	-8.6138	9.6138
High	Low	-10.33333	3.82213	.022	-19.4471	-1.2195
	Medium	-.50000	3.82213	.991	-9.6138	8.6138

Table 5: One-way ANOVA F-test output

CONTINUED

Appendix D: Pharmacy Data

Questions 34 to **43** refer to the information in this appendix.

In New Zealand about 3500 people work in community pharmacies and a register of all pharmacists working in New Zealand is maintained. A random sample of pharmacists were asked (among other things) about their professional quality of life and several measures were recorded. One such measure was called a **Compassion satisfaction** score. This was a measure (from 0 to 50) of the pleasure derived from being able to do their job well. Higher scores represent greater satisfaction.

Questions 34 and **35** refer to the following additional information.

In the sample there were 109 pharmacists who worked in a city. A bootstrap confidence interval was constructed to estimate the difference between the mean compassion satisfaction score for females who worked in a city and that for males. The three plots from the iNZightVIT bootstrap confidence interval construction output are shown in Figure 8.

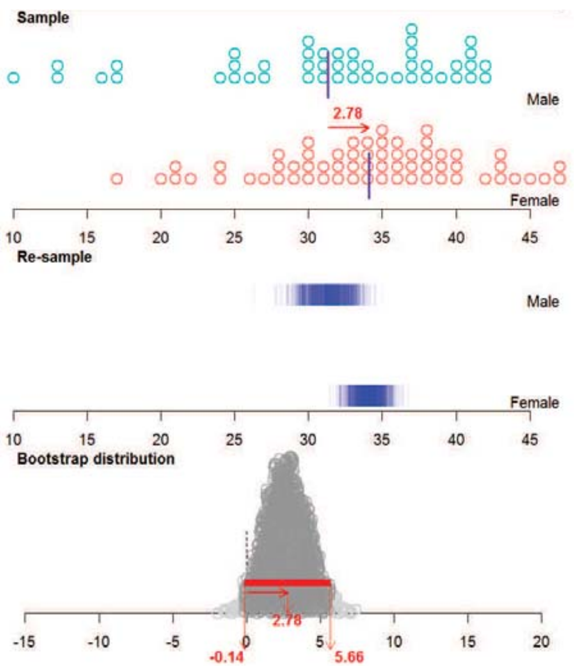

Figure 8: Bootstrap confidence interval construction output

Questions 36 to **40** refer to the following additional information.

A two-tailed two independent sample *t*-test was conducted to see whether there was a location effect between city and rural pharmacies. Computer output is shown in Figure 9 and Table 6.

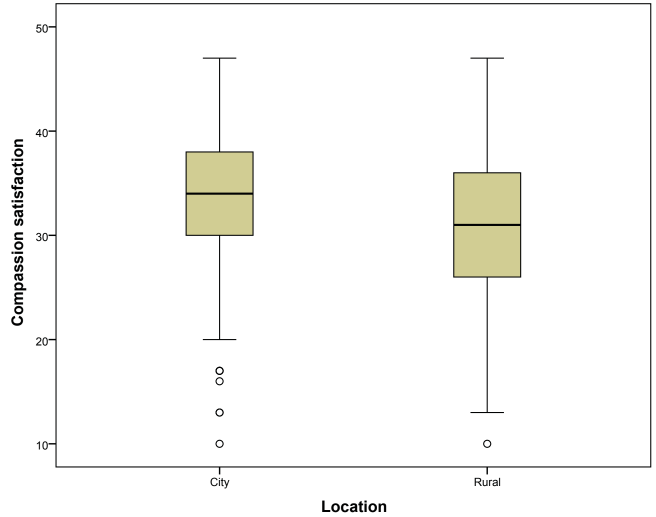

Figure 9: Compassion satisfaction scores

Group Statistics

	Location	N	Mean	Std. Deviation	Std. Error Mean
Compassion satisfaction	City	109	32.91	7.577	.726
	Rural	142	30.99	7.537	.632

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Compassion satisfaction	Equal variances assumed	.051	.822		249	.048	1.915	.962	.021	3.810
	Equal variances not assumed			++	231.88	.048	1.915	.963	.019	3.812

Note: One value has been replaced with ++.

Table 6: *t*-test output

Appendix E: NFL Data

Questions 46 to **49** refer to the information in this appendix.

Researchers were interested in finding whether there really is a home game advantage effect on the results of the National Football League (NFL) in the United States. I.e., whether teams are, on average, more likely to win playing at home than playing away. Data collected, included recording the home and away team scores for each of 256 games. Assume these games form a random sample of all NFL games.
Data sourced from Lock *et al.* (2013).

A one-tailed paired-data t-test was conducted. Box plots are shown in Figures 10 and 11 and test output is shown in Table 7, page 36.

Let:
Home be the score for the team playing at home,
Away be the score for the team playing away
and
Diff = **Home** – **Away**.

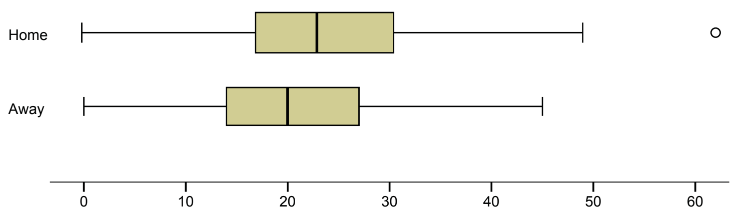

Figure 10: NFL scores for home and away games

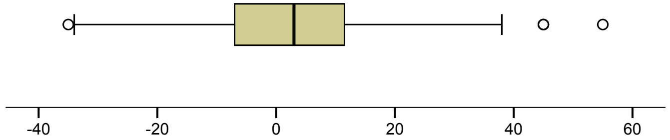

Figure 11: Differences in NFL game scores (**Home** – **Away**)

CONTINUED

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Home	23.81	256	10.524	.658
	Away	20.55	256	9.635	.602

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Home - Away	3.266	15.203	.950	1.394	5.137	3.437	255	.001

Table 7: Paired-data t-test output

ANSWERS:

1. (1)	2. (5)	3. (2)	4. (3)	5. (5)
6. (1)	7. (4)	8. (4)	9. (4)	10. (5)
11. (1)	12. (2)	13. (3)	14. (5)	15. (2)
16. (1)	17. (3)	18. (2)	19. (2)	20. (5)
21. (3)	22. (5)	23. (1)	24. (4)	25. (1)
26. (3)	27. (1)	28. (4)	29. (5)	30. (2)
31. (5)	32. (4)	33. (2)	34. (1)	35. (4)
36. (3)	37. (4)	38. (3)	39. (2)	40. (2)
41. (1)	43. (3)	44. (5)	45. (3)	46. (2)
47. (4)	48. (3)	49. (1)		