STATS 101/101G/108 Introduction to Statistics

Assignment 3, First Semester 2017

Due: 3pm Wednesday 31st May

Read these instructions carefully.

Marks

- Assignment 3 is worth 10% of your final mark**. Do not leave it until the last day**.
- It will be marked out of **85 marks**, 80 marks for the questions as shown and 5 marks for communication and presentation. See below for how these 5 marks are allocated. Your final mark will be converted to a mark out of 10 which will be recorded towards your course work.
- Statistics is about summarising, analysing and communicating information. Communication is an important part of statistics. For this reason you will be expected to write answers which clearly communicate your thoughts.
- **Communication and Presentation marks:**
  - **Demonstrated clear sentence structure**: this includes correct use of full stops and capital letters; not writing excessively long or complicated sentences; attention to spelling and grammar.
  - **Demonstrated ability to communicate information clearly in sentences**: this includes sentences easily conveying the correct idea; sentences making sense; comments not being excessively long or short; conclusions following logically from previous statements.
  - **Assignment tidily set out and easy to follow:** this includes the answers being clearly set out in the correct order; the assignment not being overly messy; graphs and plots are tidy with correct labelling of axes; the assignment including the correct cover sheet being clipped together or stapled.
  - **Follow the "Step-by-Step Guide to Performing a Hypothesis Test by Hand" as required.**

    A "*t*-test by hand" can be handwritten or typed!
  - **Student ID number shown on the assignment**: this can be on the inside of the cover sheet or on the top of the first page of the assignment.

Question guide

- Attempt questions 1, 2 and 3 when Chapter 7 has been covered.
- Attempt question 4 when the first half of Chapter 8 has been covered.
- Attempt questions 5 and 6 when all of Chapter 8 has been covered.

Hypothesis tests in this assignment

- Practical significance:
  - **Apart from question 3**, you do NOT need to interpret hypothesis tests in terms of practical significance.
- In question 4:
  - You must clearly show that you have followed steps 1, 2, 3, 7, 9 and 10 in the "Step-by-Step Guide to Performing a Hypothesis Test by Hand", Lecture Workbook, page 11, Chapter 7. The other steps are replaced by your computer output, which you **must hand in**.
- Report *P-values* to 3 or 4 decimal places.

Computer use in this assignment

- Make sure you are prepared for questions 4 and 5 before you begin to use the computer.
- Hand in all computer output for questions 4 and 5.
- When carrying out a two independent sample *t*-test using SPSS do not assume equal variances.

Notes

- The format and handing in of Assignment 3 is the same as that for Assignments 1 and 2. Refer to the instructions on page 1 of those two assignments.
- Refer to the Worked Examples file under *Assignments and Assignment Resources* on Canvas for examples of how to set out your answers.
- Refer to the Lecture Workbook, Section A (Course Information), page 3, Assignment Rules: Working together versus cheating

Question 1. [10 marks] [Chapter 7]

A psychologist was interested in whether attitudes toward death differ between organ donors (people who, on their drivers licence, indicate that they are willing to donate their organs) and non-organ donors. 25 organ donors and 69 non-organ donors were randomly selected and the extent to which each person is concerned about issues relating to death was measured using the Templar Death Anxiety Scale (DAS). The DAS produces scores ranging between 0 and 15 with higher scores indicating greater anxiety towards death. Summary statistics are displayed below:

DAS score	N	Mean	Std. Deviation
Organ donor	25	5.36	2.91
Non-organ donor	69	7.62	3.45

(a) Carry out a *t*-test to investigate whether there is a difference between the mean DAS score for all organ donors and all non-organ donors. [9 marks]

Notes:

(i) Refer to the instructions on page of this assignment: "Hypothesis tests in this assignment".

(ii) You must clearly show that you have followed the "Step-by-Step Guide to Performing a Hypothesis Test by hand" given in the Lecture workbook, page 11, Chapter 7.

(iii) At steps 5 and 8 it is necessary to use the *t*-procedures tool on Canvas to determine the standard error and the *t*-multiplier. Look under: **Assignments → Assignment 3**

(iii) At step 6 it is necessary to use the *t*-procedures tool on Canvas, a graphics calculator, SPSS or Excel to determine the *P-value*.

(b) Does the confidence interval given in part (a) contain the true value of the parameter? Briefly explain. [1 mark]

Question 2. [9 marks] [Chapter 7]

In March 2015, Sport New Zealand[1] published the report 'Sport And Active Recreation In The Lives Of New Zealand Adults' which was based on the 2013/2014 Active New Zealand Survey. For this survey trained interviewers conducted face-to-face survey interviews with a nationally-representative sample of 6430 New Zealanders aged 16 or over. Assume the sample is a simple random sample of adult New Zealanders.

One question in the survey asked for the main reasons for participating in sport and active recreation. The table below shows the results from the 6430 adult New Zealanders classified by their age group.

	Age group					
Main reasons	16 – 24 (*n* = 757)	25 – 34 (*n* = 934)	35 – 49 (*n* = 1639)	50 – 64 (*n* = 1585)	65 – 74 (*n* = 869)	75 and over (*n* = 646)
Fitness and health	695	878	1490	1412	786	542
Cultural reasons	210	342	638	407	123	52
Enjoyment	704	832	1472	1382	748	439
Social reasons	507	542	803	761	424	293
Sport performance	447	347	503	325	150	52
Low cost	384	464	747	705	315	154
Convenience	299	452	706	756	366	186

(a) State the sampling situation for analysing the difference between the estimated proportion of New Zealanders aged 16 – 24 years who included 'Enjoyment' as a main reason for participating in sport and active recreation in 2013/2014 and the estimated proportion of New Zealanders aged 25 – 34 years who included 'Enjoyment' as a main reason. [1 mark]

(b) Carry out a *t*-test to investigate whether there is a difference between the proportion of all New Zealanders aged 16 – 24 years who included 'Social reasons' as a main reason for participating in sport and active recreation in 2013/2014 and the proportion of all New Zealanders aged 16 – 24 years who included 'Sport performance' as a main reason for participating in sport and active recreation in 2013/2014. [8 marks]

Notes:

(i) Refer to the instructions on page 1 of this assignment: "Hypothesis tests in this assignment".

(ii) Follow the "Step-by-Step Guide to Performing a Hypothesis Test by Hand" given in the Lecture Workbook, page 11, Chapter 7.

(iii) At steps 5 and 8 it is necessary to use the *t*-procedures tool on Canvas to determine the standard error and the *t*-multiplier. Look under: **Assignments → Assignment 3**

(iv) At step 6 it is necessary to use either the *t*-procedures tool on Canvas, a graphics calculator, SPSS, or Excel to determine the *P-value*.

[1] Sport New Zealand, 2015. Sport and Active Recreation in the Lives of New Zealand Adults. 2013/14 *Active New Zealand Survey Results. https://www.srknowledge.org.nz/researchseries/active-new-zealand-20132014/*

Question 3. [10 marks] [Chapter 7]

Read *Confidence Intervals and P-values*. This article can be found on Canvas. Look under Assignments → Assignment 3

A confectionery factory uses imported cocoa beans to make small chocolate bars. Randomly chosen chocolate bars are tasted and given a taste quality score; a numerical value ranging from 0 to 10. Based on past data the taste quality score is, on average, 9.25 for chocolate bars made from the current source of cocoa beans. It is known that cocoa beans from different sources can affect the taste quality of the chocolate bars.

Management has been advised that sales would increase if the current mean taste quality score can be increased by at least 0.3, whereas sales would decrease if the mean taste quality score drops by 0.5 or more, assuming all other factors remain fixed. Any change in the mean taste quality score of between these two values would be of no consequence with respect to sales.

A study is conducted by the quality control team to determine what effect a new source of cocoa beans will have on the current taste quality mean score of 9.25 for the purpose of identifying a sales effect.

Some possible outcomes of the study using the new source of cocoa beans are:

	$\bar{x}$	se($\bar{x}$)	*P-value*	95% CI
Case 1	9.72	0.0688	0.0000	(9.59, 9.85)
Case 2	9.31	0.1124	0.5938	(9.09, 9.53)
Case 3	8.87	0.4698	0.4190	(7.95, 9.79)
Case 4	8.17	0.1376	0.0000	(7.90, 8.44)
Case 5	9.01	0.0390	0.0000	(8.93, 9.09)

Note:

The hypotheses associated with the quoted *P-values* are:

H_0 : The mean taste quality score is 9.25.

H_1 : The mean taste quality score is not 9.25.

(a) **(i)** What is the hypothesised value? [1 mark]

(ii) In which direction and how far away, in terms of standard errors, is the estimated taste quality score in Case 1 ($\bar{x} = 9.72$) from the hypothesised value? [2 marks]

(b) Which case(s) demonstrates, at the 5% level, that the sample mean, $\bar{x}$, is significantly different to the hypothesised value? [1 mark]

(c) For which case(s) are we able to claim that the true mean taste quality score using the new source of cocoa beans:

(i) has practical significance? [1 mark]

(ii) does not have practical significance? [1 mark]

(d) In which case(s) have we learned nothing useful about the true mean taste quality score using the new source of cocoa beans? [1 mark]

(e) Suppose the actual outcome for the study is:

	$\bar{x}$	se($\bar{x}$)	*P-value*	95% CI
Case 6	8.45	0.1491	0.0000	(8.16, 8.74)

Write three to five sentences interpreting this output. You need to refer to statistical significance and practical significance. Which source of cocoa beans (current or new) would you recommend? Give a reason(s) for your choice. [3 marks]

Questions 4 and 5 refer to the following information.

A study[2] on customers of Boston (USA) coffee shops was conducted. The field study was interested in identifying or quantifying the presence of discrimination against customers in stores, restaurants and other small transaction consumer markets. Researcher assistants visited eight coffee shops and recorded information on orders made by 286 customers. Three of the variables used in the study are described below.

Variable	Type
Sex	The customer's sex: Female, Male
Age group	The age group of the customer (years): 15 to 25, 26 to 39, 40 and over
Waiting time	The time between ordering and receiving coffee (in seconds)

Note:

The sample data used in Questions 4 and 5 have been simulated and are consistent with summary statistics provided in the paper.

Question 4. [15 Marks] [First half of Chapter 8]

We wish to investigate whether the waiting times differed between female and male customers. The waiting times of 141 female customers and 145 male customers were recorded.

Notes:

(i) To answer parts (c) and (d) you need to ensure that you use the file(s) which has the data in the **form that is appropriate** for the design of the study.

(ii) SPSS and Excel filed of the data are available on Canvas on the **STATS 10x Front page** or look under **Assignments → Assignment 3**.

Click on:

- `WaitingTimeData–A–iNZight` or `WaitingTimeData–A–SPSS`
- `WaitingTimeData–B–iNZight` or `WaitingTimeData–B–SPSS`

(a) What type of study is this: Experiment or Observational study? Briefly justify your choice. [2 marks]

(b) For this study describe the: [1 mark]

(i) units,

(ii) treatment or factor of interest,

(iii) response.

(c) **(i)** Using **iNZight**, draw the appropriate plot(s) for this data set. (You should consider the design of this study to ensure the relevant plot(s) is drawn.) Do not use SPSS to draw the plot(s). [1 mark]

(ii) Comment on any features in the plot(s). [3 marks]

(d) Investigate whether, on average, there is a difference between the waiting times of female customers and those of male customers. Use **SPSS** to conduct a *t*-test. Interpret your results. (You should consider the design of this study to ensure the appropriate *t*-test is conducted.) [6 marks]

Reminder: Refer to the instructions on page 1 of this assignment: Hypothesis tests in this assignment.

(e) Comment on the validity of the *t*-procedures conducted in (**d**) by briefly discussing each assumption. [2 marks]

[2] Myers, C., Bellows, M., Fakhoury, H., Hale, D., Hall, A., and Ofman, K. (2010). Ladies first? A field study of discrimination in coffee shops. *Applied Economics*,44(2), 142–147.

Question 5. [21 marks] [Second half of Chapter 8]

Of interest was whether the waiting time differed depending on the age group of the customer. The 286 customers were categorised into three age groups: 15 – 25, 26 – 39, 40 and over and their waiting times were recorded.

Note:

SPSS and Excel files of the data are available on Canvas on the **STATS 10x Front page** or look under **Assignments → Assignment 3**.

Click on:

- `CoffeeShopData-SPSS`
- `CoffeeShopData-iNZight`

(a) **(i)** Use **iNZight** to draw the appropriate plots(s) for this data set. [1 mark]

(ii) Comment on any features in the plot(s) in terms of the original story. [4 marks]

(b) Using SPSS provide the computer output of an *F*-test on these data.

Notes:

- Refer to the SPSS Tutorial, pages 16 and 17, on Canvas. (Look under Software Information and Help → SPSS Help.)
- Ensure that you complete Step 1 through to Step 4 of the instructions on pages 16 and 17.

(c) State the assumptions of the *F*-test in terms of the original story. [4 marks]

(d) Calculate the ratio of the largest sample standard deviation for the waiting times to the smallest sample standard deviation for the waiting times. [1 mark]

(e) Comment on the validity of the *F*-test by briefly discussing each assumption. [3 marks]

(f) Assume that an *F*-test is an appropriate test to use here. (Note: It may not be.)

(i) State the null hypothesis for the test, both in words and using symbols. [1 mark]

(ii) State the alternative hypothesis for the test in words. [1 mark]

(iii) What does the result of the *F*-test tell you about the underlying mean waiting times for customers in the three age groups? Explain your answer in 1 to 2 sentences. [1 mark]

(g) **(i)** Assuming the Tukey's pairwise comparisons are valid and appropriate.
Investigate whether, on average, there is a difference between the waiting times for customers aged 15 to 25 years and that for customers aged 40 and over years. Interpret the *P-value* and confidence interval. [2 marks]
Note: A conclusion is <u>not</u> required here.

(ii) Between which pair (or pairs) of age groups were there significant differences (at the 5% level) in the mean waiting times? [1 mark]

(iii) Are we able to determine which age group has the longest underlying mean waiting times? If so, name the age group. [1 mark]

(h) In one to three sentences, provide an overall conclusion for this study. [1 mark]

Question 6. [15 marks] [Chapters 7 and 8]

Below is some information regarding variables obtained from a survey of over 800 banking customers across New Zealand.

Variable	Type
Satisfaction	The customer's overall satisfaction with their main bank (on a scale from 0 – 10)
Closeness	The customer's perceived closeness of the relationship with their personal banker (on a scale from 0 – 10)
Bank	The customer's main bank: ANZ, ASB, BNZ, Kiwibank, Westpac, Other
OnlyBank	The customer's main bank is their only bank: Yes, No
Sex	The customer's sex: Male, Female
Product	The main banking product used by the customer: online account, current account, savings account, investment, loan
Income	The customer's personal yearly income (in thousands of dollars)
Performance	The customer's rating of the overall level of performance for their main bank (on a scale from 0 – 10)
Advice	The customer's rating of availability of financial advice from their main bank (on a scale from 0 – 10)

(a) For each of the scenarios **1** to **5** below: [5 marks – 1 mark for each scenario]

(i) Write down the name of the variable(s), given in the table above, needed to examine the question.

(ii) For each variable in (i) write down its type (numeric or categorical).

(b) What tool(s) should you use to begin to investigate the scenarios **1** to **5** below? Write down the scenario number **1** to **5** followed by the appropriate tool. **Hint**: Refer to the notes in Chapter 1 in the Lecture Workbook. [5 marks – 1 mark for each scenario]

(c) Given that the underlying assumptions are satisfied, which form of analysis below should be used in the investigation of each of the scenarios **1** to **5** below? Write down the scenario number **1** to **5** followed by the appropriate Code **A** to **F**. (See Page 5).
[5 marks – 1 mark for each scenario]

Scenario 1 Is there is a difference between the personal yearly incomes of customers with one bank and that of customers with more than one bank?

Scenario 2 Is a customer's overall satisfaction with their main bank related to their main bank?

Scenario 3 Does a customer's rating of overall performance of their main bank differ from their rating of perceived closeness of the relationship with their personal banker?

Scenario 4 Is there a difference between the proportion of female customers with one bank and the proportion of male customers with one bank?

Scenario 5 Does a customer's rating of availability of financial advice from their main bank depend on the main banking product used by the customer?

Code	Form of analysis
A	One sample *t*-test on a mean
B	One sample *t*-test on a proportion
C	One sample *t*-test on a mean of differences
D	Two sample t-test on a difference between two means
E	Two sample *t*-test on a difference between two proportions
F	One-way analysis of variance *F*-test