# STA10003 FOUNDATIONS OF STATISTICS ASSIGNMENT PART 1

This **Assignment - *Part 1*** is worth **20%** of your final mark for STA10003.

## The Industry Scenario

You are a new graduate researcher at a social science and psychological sciences research institute, and you have been given a dataset to analyze which relates to a study of Californian adults in 2020. The survey, titled the California Health Interview Survey [CHIS] collects extensive information regarding health status, health conditions, health related behaviours, health insurance coverage and other health-related issues and demographic information. You have been tasked with conducting the initial analysis of some variables using R and to write brief reports.

## Data Preparation

For your Assignment Part 1 you should use the SPSS data file **STA10003 Sem 1 2023 Assignment Data.sav**.

You must use R to *draw a* ***random sample of 1500*** *of the 6259 observations in the* **STA10003 Sem 1 2023 Assignment Data.sav** data file. You will conduct your analysis using this random sample of 1500.

## Submission Instructions

* Your submission must be a single Word file.
* Although a cover page is not required, you should include your name and student number within the document [eg in footer].
* You must submit your file via Canvas by the specified due date and time to avoid late penalties. Only the last document you submit will be retained by Canvas.
* Once submitted, please review your submission to ensure the correct file has been submitted.
* This is an individual assignment. Do not share your work with other students. They will have a different random sample of data, so any copying will be detected.

**ASSIGNMENT – *PART 1***

### For your Assignment - *Part 1*, you are required to complete the first three [3] questions by producing the appropriate analyses and writing the relevant report for each question. You are also required to complete questions 4 and 5, which contain short answer questions.

***For each of the first three questions requiring R, you should include the relevant output [eg tables and graphs] with your report. Note that many of the variables have similar names, so it is important that the correct variable be selected to address the question asked.***

***Question 1*: Self-reported Health Status**

The variable **HealthStatus** indicates the self-reported health status of each of the Californian adults. Using R, summarize the **HealthStatus** variable and write a paragraph explaining the key features of the data observed in the output in the report writing style.

***Question 2*: Cigarette Consumption**

The variable **Cigarettes** gives an indication of number of cigarettes each adult reported smoking in the previous day. Using R, summarize the **Cigarettes** variable and write a paragraph explaining the key features of the data observed in the output in the report writing style.

Graphical user interface, text, application

Description automatically generated

***Question 3*: Walking for Leisure**

The variable **WalkLeisure** gives an indication of the time each participant spent walking for leisure. This variable was measured in minutes spent walking in the previous week. Using R, produce the relevant output to summarize the **WalkLeisure** variable and write a paragraph explaining the key features of the data observed in the output in the report writing style.

***Question 4*: [does not require R]**

Dale is a psychology student who enters three competitions in which puzzles need to be solved as quickly as possible by correctly fitting all the pieces of a gadget together. In Competition A, Dale takes 80 seconds to complete the puzzle, in Competition B he takes 84 seconds, and in Competition C he takes 85 seconds to complete the puzzle.

Overall completion times for participants in Competition A are normally distributed with a mean of ** = 85 seconds and a standard deviation ** = 8 seconds. Completion times for participants in Competition B are normally distributed with a mean ** = 90 seconds and a standard deviation ** = 5 seconds. Completion times for participants in Competition C are normally distributed with a mean ** = 80 seconds and a standard deviation ** = 3 seconds.

In which competition was Dale’s performance better, relative to others who took part in that competition? Justify your answer, quoting relevant statistics as part of your explanation.

### *Question 5*: [Does not require R. No calculations are required]

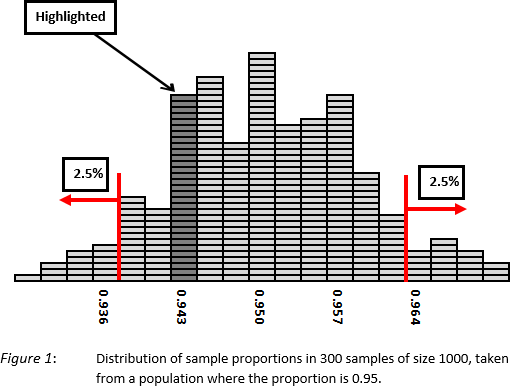
Australian Consumer Law has mandatory packaging standards relating to biodegradable / compostable products. The Bioware Company manufactures coffee cups and lids, and in order to display a progressive impression to consumers, they have ceased producing plastic coffee cup lids and have replaced these with the production of compostable lids. All coffee cup lids require a hole punched in them for the transfer of air. Despite the use of very accurate equipment in the production of coffee cup lids, there are always some lids in each batch that do not meet the Australian standards, where dents rather than holes are pierced in the lids.

The equipment at the Bioware factory has recently been replaced and the production line manager wants to check that it is performing well. If less than 95% of the compostable coffee cup lids produced meet the Australian standards, the equipment will need to be recalibrated, which involves a lengthy and expensive process. The manager takes a random sample of 1000 compostable coffee cup lids produced during July to investigate the proportion of lids that meet the Australian standards.

1. What is the population we can draw conclusions about in this study?

* All compostable coffee cup lids produced by the Bioware Company during July.

To answer questions (b) to (d), consider **only** the sampling distribution shown in *Figure 1*.



1. What does the highlighted section of the distribution in *Figure 1* represent?

* All samples of 1000 compostable coffee cup lids with a proportion near 0.943.

1. The random sample of 1000 lids taken by the production line manager turned out to have a proportion of 0.929. Does this sample look like it belongs to the sampling distribution displayed in *Figure 1*? Justify your answer.

* No, the sample does not look like it belongs to the sampling distribution in Figure 1. The proportion is more than

1. Given that the sample was randomly selected from all compostable coffee cup lids manufactured by the Bioware Company in July, and given that the sample proportion was measured accurately, what conclusion can we reach from part (c)?

### *Note:* Your data file [containing a random sample of 1500 cases] will also be used in Assignment [*Part 2*], so ensure that you save it in a ‘safe’ place.