# STAT1301 Advanced Analysis of Scientific Data

## Assignment 4

The due date/time is given on Blackboard. For auto-marked questions, please enter your final answer on gradescope. No working is required. For the written question, please upload your working as a single pdf file.

**Auto-marked questions** *For all questions in this section, please give your answer to 2 decimal places rounding normally and use the course tables for probability calculations.*

1. A hospital is trialling a new protocol intended to lower LDL-cholesterol (a cardiovascular risk biomarker). A simple random sample of 10 patients following the protocol for 12 weeks produced a sample mean change in LDL of 2.70 (mmol/L) with sample variance 0.25 (mmol/L).

   Based on this data, calculate the margin of error for a 90% confidence interval.

2. A workplace ergonomics study compared typing speeds (in words per minute, WPM) of employees at two types of workstations:

   - Standing desk: $n_1$ = 10, sample mean = 50.2, sample variance = 10.84 (WPM$^2$)
   - Stepping desk: $n_2$ = 20, sample mean = 38.5, sample variance = 5.25 (WPM$^2$)

   The company claims that employees at stepping desks type, on average, 10 WPM slower than those at standing desks. The company says that there is no reason to suspect the variation is the same for the two different desk types.

   (a) You believe their claim may be inaccurate and want to test whether the stepping desk has a population mean that is worse than the standing desk by more than 10 WPM. What is the correct set of hypotheses?

   (b) Calculate the associated test statistic for the hypothesis test. Report your answer using 2 decimal places.

   (c) Based on your test statistic and hypotheses, find the associated p-value using the tables. Report the lower and upper bound given by the tables - i.e. Lower < p-value < Upper.

3. A study was conducted to examine the relationship between technology use and sleep quality.

   They randomly sampled 40 university students who reported using screens for at least two hours before bed each night, and 35 students who reported using no screen time before bed. For each participant, the number of hours of sleep on a typical weeknight was recorded.

   The study aims to determine whether there is a significant difference in mean sleep hours between the two groups. You can find the associated dataset on blackboard. For this dataset you can load the data using:

   *Sleep = read.csv("sleep.csv")*

   (a) Based on the dataset provided what is your estimate for the population mean number of hours of sleep for students who report no screen time before bed? (Report to 2 decimal places).

(b) Calculate the pooled variance of the sleep hours across the two groups. (Report to 2 decimal points)

(c) Calculate the p-value for a two sample test (assuming that the population variances are equal for the two samples) with the hypotheses:

$$H_0 : \mu_{\text{no screen}} - \mu_{\text{high screen}} = 0 \qquad \text{vs} \qquad H_1 : \mu_{\text{no screen}} - \mu_{\text{high screen}} \neq 0$$

Based on this p-value, what is the *most* appropriate conclusion using a significance level of 5%.

## Written questions

4. Answer questions a to f using the following information.

A study was conducted to evaluate whether mailing HPV self-sampling kits improves cervical cancer screening uptake compared with education-only outreach. A random sample of 1,415 patients who were overdue for screening received a direct-mail self-sampling kit, of whom 505 completed screening. An independent random sample of 1,408 comparable patients received education-only outreach, of whom 264 completed screening. The investigators hypothesised that the proportion completing screening would be higher under the direct-mail strategy than under education alone.

(a) State the relevant statistical hypotheses required to answer the researcher's question in mathematical form. Explain all notation used. [1 mark]

(b) Let $X$ be the number of patients who received the direct mail strategy and completed screening and let $Y$ be the number of patients who received the education only outreach and completed screening. State the distributions for $X$ and $Y$, state the normal approximations for the sample proportions, and comment on if the aproximations are reasonable for this research problem. [1 mark]

(c) Perform the appropriate hypothesis test. Show all working and report the relevant statistic to three significant figures. Make a statistical conclusion and infer what this means in the context of the study. [3 mark]

(d) Produce a 97% confidence interval for the difference between the two strategies in terms of the proportion who completed screening. Using the 97% confidence interval, comment on if there is evidence that the proportions differ. [1 mark]

Consider the following setting of a two sample proportion test. Let $X \sim Bin(n_X, p_X)$ and $Y \sim Bin(n_Y, p_Y)$ be independent random variables corresponding to the number of observations recorded for the two populations. You decide to consider an estimate $\hat{P}^W = w\frac{X}{n_X} + (1-w)\frac{Y}{n_Y}$ which is a weighted average of the two sample proportion estimates.

(e) Under the null hypothesis (that $H_0 : p_X = p_Y = p$), show that $\mathbb{E}[\hat{P}^W] = p$ and the variance of $\hat{P}^W$ is

$$\mathbb{V}\text{ar}(\hat{P}^W) = p(1-p)\left(\frac{w^2}{n_X} + \frac{(1-w)^2}{n_Y}\right)$$

[2 mark]

(f) Show the variance is minimised when $w = \frac{n_X}{n_Y + n_X}$. Note that this corresponds to $\hat{P}^W = \frac{X+Y}{n_X+n_Y}$, which is the pooled proportion estimator. [2 mark]