EDGE stats lab

The goal of this assignment is to increase your understanding of the key statistical concepts of population, sample, standard deviation, significance, power and effect size. It does this in the context of the two sample t-test. Through completion of this lab, it is expected you will become familiar with the EDGE (Excel Dataset GEnerator) spreadsheet (available at the EDGE website linked via Learn). The spreadsheet allows you to carry out experiments that will help you increase your understanding of these key statistical concepts.

1. **Generating a sample under a given model, carrying out t-tests and calculating confidence intervals.**

Scenario: Exam scores for all Loughborough University graduates who studied Psychology or Geography were gathered. For these populations, it was found that the scores were normally distributed with a mean score for Psychology students of 62%, and a mean score for Geography students of 56%. The standard deviation for Psychology students was 4%, and the standard deviation for Geography students was 5%. In this question, we will examine how to use data collected from a SRS from each of these populations to make inferences about these population parameters if we did not know them.

(a) Open the spreadsheet entitled 'Two sample t-test'. Use this tool to generate exam scores for a SRS of size 15 from each population, Psychology students and Geography students, using the instructions at the top of the spreadsheet. List the values you generated on your answer sheet.

(b) We are interested in determining from our sample if exam scores for Psychology and Geography students are different. Determine the null and alternative hypotheses relevant to this question. Which one is true given the way you have chosen your data?

(c) Calculate the mean values of sample 1 and sample 2, $\hat{y}_1$ and $\hat{y}_2$, by hand. Check these values with those calculated by the spreadsheet in the green cells labelled $\hat{y}$.

(d) Calculate the t-statistic by hand using the sample statistics produced in the green box. Check this against the result that is produced in the spreadsheet labelled 't-statistic'.

(e) Calculate the degrees of freedom by hand using the sample statistics produced in the green box. Check this against the result that is produced in the spreadsheet labelled 'df by hand'.

(f) Find the critical t-value using your degrees of freedom at the 0.05 significance level. There is a t-distribution table located in the spreadsheet for you to

obtain your critical t-value. The value that you find may differ slightly from the one produced using the software.

i. What can you conclude about the significance of your t-statistic?

ii. What does this mean in terms of your hypotheses?

(g) Use your degrees of freedom, t-statistic and the t-distribution table to approximate the p-value.

i. Compare your p-value estimate with the precise p-value calculated in the green cell labelled 'p-value by hand'.

ii. Comment on the significance of this precise p-value at a 95% confidence level.

(h) Calculate a 95% confidence interval around the difference in means for your sample by hand. Check your answer with the confidence interval produced in the spreadsheet labelled '95% C.I. around difference in means by hand'.

i. Is 0 included in the interval?

ii. How is this related to the p-value that you estimated?

2. **Interpreting p-values and confidence intervals.**

Set the population mean for Psychology students, μ_1, to equal the population mean for Geography students, μ_2, i.e. $H_0 : \mu_1 - \mu_2 = 0$. Delete your sample from question 1 by highlighting the first purple column, right clicking and selecting 'Clear Contents'. Set μ_1 and μ_2 both equal to 60 and set σ_1 and σ_2 both equal to 5. Using the instructions in the spreadsheet, generate 20 datasets in the purple columns consisting of two samples, each of size 15. The p-values for each dataset will automatically be recorded in the yellow column labelled 'P-value' near the bottom of the spreadsheet. The goal of this question is to investigate the relationship between p-values, significance level and confidence intervals.

(a) Insert a screenshot of the spreadsheet page here.

(b) Is the null hypothesis of the two-sample t-test true or false for these samples?

(c) Define p-value

(d) Define significance level

(e) How many of your p-values are significant at a 95% confidence level? How many do you expect to be significant from the definition of confidence level?

(f) How many p-values would you expect to be significant at the 95% confidence level from 100 samples? How many would you expect to be significant at the 99% confidence level from 100 samples?

(g) The spreadsheet will automatically plot a histogram of p-values. Explain how this compares to your expectations for the histogram from the definition of confidence level.

(h) The spreadsheet will automatically plot the p-values in a scatter plot. Is there any pattern? Discuss what you observe.

(i) Define confidence interval.

(j) The spreadsheet will automatically record the value for $\hat{y}_1 - \hat{y}_2$, the lower bound of the confidence interval and the upper bound of the confidence interval in the yellow box at the bottom of the spreadsheet. Similarly, the spreadsheet will automatically plot the confidence intervals on a graph in the spreadsheet. How many of the confidence intervals include 0? How does this relate to the p-values? Discuss what you observe.

3. **Experimenting with significance, sample size, standard deviation, effect size and power.**

Recall from the lecture notes:
When we make decisions in statistics, there are four possible scenarios. We can arrange these in a table:

	H_0 is true ($\mu_1 = \mu_2$)	**H_a is true** ($\mu_1 \neq \mu_2$)
Significant (Test rejects H_0)	**Type I error**	Correct
Non-significant (Test does not reject H_0)	Correct	**Type II error**

(a) In Question 2, which part of the table above was relevant?

(b) Clear your datasets from question 2 by highlighting the purple columns, right clicking and selecting 'Clear Contents'. Enter your values from question 1 into the blue cells: $\mu_1 = 62$, $\mu_2 = 56$, $\sigma_1 = 4$ and $\sigma_2 = 5$ and generate 20 datasets in the purple columns consisting of two samples, each of size 15. For each dataset, the p-value, the value for $\hat{y}_1 - \hat{y}_2$ and the lower and upper bounds of the 95% confidence interval will automatically be recorded in the yellow boxes. Take screenshots of the yellow boxes and graphs, and paste these into a Word document as you will need to compare these values with those produced in the following parts.

(c) For this data, which part of the table above is relevant? Why?

(d) Define the power of a statistical test.

(e) Now using the results in the spreadsheet, estimate the power of the t-test here against the effect size that we have: $\epsilon = \mu_1 - \mu_2 = 6$. Also estimate the width of the confidence intervals as the average width of the confidence intervals you have obtained.

(f) Clear the contents of the purple columns. Increase the **sample size** to 50 and generate 20 new datasets consisting of two samples, each of size 25. The p-values, the value for $\hat{y}_1 - \hat{y}_2$ and the lower and upper bounds of the 95% confidence interval will be recorded in the yellow boxes. Again, take screenshots of the yellow boxes and graphs, and paste these into the Word document. Now from the new results, estimate the power and the width of the confidence intervals. In the pink boxes on the spreadsheet state the change from part (a) in the following.

   i. Power estimate.

   ii. Width of the confidence interval.

(g) Clear the contents of the purple columns. Reset the sample size to 30, increase the **standard deviation** by setting $\sigma_1 = 10$ and $\sigma_2 = 12$, and

generate 20 new datasets consisting of two samples, each of size 15. The spreadsheet will record the p-values, the value for $\hat{y}_1 - \hat{y}_2$ and the lower and upper bounds of the 95% confidence interval in the yellow boxes. Again, take screenshots of the yellow boxes and graphs, and paste these into the Word document. Estimate the power and width of confidence intervals in this situation. In the pink boxes on the spreadsheet state the change from part (a) in the following.

i. Power estimate.

ii. Width of the confidence interval.

(h) Clear the contents of the purple columns. Reset σ_1 to 4 and σ_2 to 5, increase the **effect size** to 12 by keeping $\mu_1 = 62$ and setting $\mu_2 = 50$, and generate 20 new datasets consisting of two samples, each of size 15. The p-values, the value for $\hat{y}_1 - \hat{y}_2$ and the lower and upper bounds of the 95% confidence interval will be recorded in the yellow boxes. Again, take a screenshot of the yellow boxes and paste into the Word document. Estimate the power and width of confidence intervals in this situation. In the pink boxes on the spreadsheet state the change from part (a) in the following.

i. Power estimate.

ii. Width of the confidence interval.

(i) For each of the cases in (a), (b), (c) and (d) use the Russ Lenth applet on the link below to calculate the power. Record the power you obtain from the applet for each case.
`http://homepage.stat.uiowa.edu/~rlenth/Power/`
Do these results agree with your answers in the pink boxes for the change in power?

4. Watch the 'dance of the p-values' video:
`http://www.youtube.com/watch?v=ez4DgdurRPg`

(a) What does it mean to say that a measurement is reliable?

(b) Are p-values on a sample a reliable measure of the difference between two populations? Why or why not?

(c) What advantages do confidence intervals have in this regard over p-values?