MACQUARIE University
SYDNEY·AUSTRALIA

STAT2170 AND STAT6180
APPLIED STATISTICS

SGTA Week 3 : Two sample testing equality of means; Sample Size and Power 2022

Part 1

For each of the following three experiments, use **R/RStudio** to carry out the appropriate statistical test. The 5% level of significance should be used.

The data sets are available for download from the unit's iLearn webpage. In order to carry out each of these tests, you will need to firstly determine:

- the appropriate test (two sample t or paired t test), including a very short statement justifying the statistical test chosen;
- whether you wish to carry out a *one* tailed or a *two* tailed test;

Then, produce and comment on relevant data summary (mean(s), std(s), etc), and graphical displays (histograms, boxplots, Normal Q-Q plots, ect.). Then perform the test. In each test:

- Define any necessary **population parameters**
- State the **null** and **alternate hypotheses** in terms of these parameters
- **Check** and **discuss** relevant assumptions for the test chosen
- Show the **test statistic** and its related **null distribution**
- Give a brief **statistical and contextual conclusion** including an appropriate confidence interval.

Question 1:

The effect of the light from two different types of 60 watt soft-white light globes was tested by measuring the lumen output of the two brands. Twenty two light bulbs (11 of each brand) were randomly selected from the total output from a week's production and the lumen output for each measured. The data, `light.dat`, are:

Brand1	75	82	101	92	79	63	103	73	87	68	60
Brand2	85	78	63	50	69	45	57	82	72	71	69

- Test if the lumen output is different between Brand1 and Brand2.

Question 2

The effectiveness of spraying to reduce the effects of corn borers in corn was evaluated on 14 farms. On each farm, two equally sized plots were selected, and then one was sprayed but the other plot was left unsprayed. The yield (in bushels per acre) was recorded. Test if the yield increases for the sprayed plots. The data, `farm.dat`, are:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
sprayed	65.3	78.1	93.0	80.7	89.0	79.9	90.6	102.4	64.7	106.1	107.4	74.0	64.6	69.5
unsprayed	70.0	74.4	86.6	79.2	84.7	75.1	87.3	98.8	70.2	101.1	97.4	65.2	68.1	68.4

Question 3

The following data is the percentage of the fine gravel component obtained from two random samples of soils. One sample was selected from relatively fertile area and the other sample from non-fertile area in a National Park. The aim of the experiment was to determine whether there was a difference in the percentage fine gravel between the two levels of soil fertility. The data (percentages of the fine gravel), `fertile.dat`, are:

fertile	5.9	8.8	6.5	13.8	20.2	16.1	7.6
nonfertile	10.6	5.4	6.1	8.2	11.5	9.1	6.7

Question 4 (to be done by hand)

A drug manufacturer has developed an oral pain medication (new drug) that they claim is absorbed into the body over a much longer period than a standard pain medication (thereby having longer pain relief). With the standard medication, the average time to reach the maximum blood concentration of the drug (after taking it) is 126 minutes with a standard deviation of 10.7 minutes. In order to validate their claim the Therapeutic Goods Administration (TGA) requires the company to use a 1% level of significance. Further, the TGA will not recognise differences of less than 5 minutes (ie, the minimal difference of interest to TGA).

a. The drug manufacturer needs to run a clinical trial (experiment on patients/people) to test their claim on the new drug at 1% significance level required by the TGA. Assume the standard deviation of the time to the maximum blood concentration of the new drug is the same as that for the standard drug, ie, 10.7 minutes
   - What are the null and alternative hypotheses for the test?
   - How many people should they enrol in their trial?
   - Without doing further calculation, explain if you need larger or smaller sample size if the significance level to be used is 5%.

b. Suppose the drug company decided to run the trial with 30 patients. What would the power of the test be if the true mean of the new drug is
   - 132?
   - 135?
   - 130?

Part 2: More on Working Directory

A working directory is a folder/location on your computer's hard drive where `R` will look for files you want to load and where it will put any files you save. The exercises below will explore the concept of working directory a bit further.

a) Open a new `R` session and find the current working directory with `getwd()`.
   - Hint: the full path of a working directory from a window machine could be something like `C:/Users/hanse/Desktop`; a working directory on a Mac machine could be something like `/Users/hanse/Documents/`

b) You can set your working directory to a different folder by using `setwd()`. Correct all the errors in the codes below, which try to set the working directory to be `C:/Users` on a windows machine or `/Users` with a Mac.

```
# For windows machine:
setwd("C:\Users")
# For a Mac
setwd("\Users")
```

c) Set your desktop as your working directory in `R` by using `setwd()`

- Hint: the code that sets my window machine as working directory is

```
setwd("C:/Users/hanse/Desktop")
```

d) Check if your working directory has been set to your desktop by using `getwd()`.

e) Download the SGTA dataset `light.dat` to your desktop from our iLearn space.

f) Correct all the errors in the codes below, which try to import `light.dat` to `R` and name it `light` after setting up the working directory":

```
light = read.table(light.dat, header = TURE)
```

g) Correct all the errors in the codes below, which try to calculate the standard deviations of the two variables in `light`.

```
sd(light$Brand1), sd(light$Brand2)
```

- One can use `t.test()` to run a two-sample t-test (with equal variance assumption) between the two variables in `light`:

```
out = t.test(light$Brand1, light$Brand2, var.equal = TRUE)
names(out)
```

  Can you extract the p-value from the `out` object directly?

- Close your `RStudio` (or `R`) and open it again. Now check if your working directory is still set to your desktop.

Problem of using `setwd()`

In the example above, the directory of the desktop depends on the username of the computer and which operating system you are using. Since this username may differ in another computer, this directory may no longer work when one shifts to another computer.

The chance of the `setwd()` command having the desired effect – making the file paths work – for anyone besides its author is 0%. It's also unlikely to work for the author one or two years or computers from now.

To recreate and perhaps extend the code, the lucky recipient will need to hand-edit one or more paths to reflect where the project has landed on their machine.

When you have to do this for the 73rd time in 2 days, this drive Jenny Bryan, statistician and software engineer at `RStudio` to post the following on `twitter`

> *If the first line of your R script is*
>
> ```
> setwd("C:\Users\jenny\path\that\only\I\have")
> ```
>
> *I will come into your office and SET YOUR COMPUTER ON FIRE* 🔥.

Solution: We can instead use the `project` in `RStudio`, which is preferable and will be discussed in detail next week.