# MSO2300 Data analysis assignment 2

Nick Sharples

2023-1-26

## Introduction

### Deadline:

Friday 10th March at 23:59

#### Learning outcomes

- **Knowledge 1 -** demonstrate significant judgement in summarising datasets using appropriate statistics and interpreting these values in context
- **Knowledge 3 -** use statistical theory to justify claims about data and identify common errors in statistical reasoning
- **Skills 4 -** perform regressions and interpret the goodness-of-fit of these models with reference to ANOVA and ANCOVA
- **Skills 6 -** demonstrate advanced skill in visualising and summarising datasets in R.

#### Marking

The **Assessment Criteria** can be found on pages 18 - 21 of the Module Handbook.

| Component | Marks |
|---|---|
| Mathematics, statistics and data-analysis | 40 marks |
| Presentation | 5 marks |

This work must **not** be done in a collaborative environment. You can use the rstudio server at rstudio.mdx.ac.uk but you must **not** do this work in the shared MSO2300/3311 project.

Instead you should either

- create your own project (Choose "New Project' ' in the Project Menu) or
- work in the default project (Choose "Close Project' ' in the Project Menu)

Work should be submitted as a .pdf for handwritten work. Any R code should be submitted as either a .R file with comments, or a .Rmd (R Markdown) file.

# Questions

## Question 1: Variance hypothesis test (Total 8 marks)

A client wishes to compare two investments:

- Apple stocks, a publicly traded company, and
- SecretFund, a private equity fund.

There is lots of data available for Apple (as it is publicly traded). The log-returns for the last 30 days are as follows:

```
##  [1]  0.0007649317 -0.0032724276  0.0266994801  0.0146942048  0.0191529700
##  [6] -0.0198222539 -0.0309390842  0.0728344816 -0.0155303302 -0.0176984263
## [11] -0.0380186151 -0.0433303238 -0.0036067267  0.0038946533  0.0041663878
## [16] -0.0337532868  0.0852364933  0.0190854637 -0.0095308998  0.0117995070
## [21] -0.0083660098  0.0128879487  0.0037746610 -0.0219186732  0.0145547228
## [26]  0.0059088238 -0.0197881042 -0.0266153293 -0.0213750947  0.0474501254
```

This data is also available in the file "Apple_Returns.csv".

A **return** $r$ is the ratio of an investment's value from one time to another. e.g.

$$r = \frac{\text{today's value}}{\text{yesterday's value}}$$

so $r > 1$ would mean the value has increased, and $r < 1$ would mean the value has decreased.

The above data are the **log-returns**, i.e. $\log r$. The log-returns are more interesting to financial analysts as their distributions are easier to understand.

We assume that the above Apple log-returns are **normally distributed**, and that they are independent and identically distributed.

Over the same time-period you have the following log-returns from SecretFund:

```
##  [1] -0.025058152  0.007345733 -0.033425144  0.063811232  0.013180311
##  [6] -0.032818735  0.019497162  0.029532988  0.023031254 -0.012215535
```

This data is also available in the file M001 Q1 data.csv

We also assume that the above SecretFund log-returns are **normally distributed**, and that they are independent and identically distributed. This might not be the same distribution as the Apple log-returns, however.

You need to advise a client about which of these funds has a higher variance (which is one way of measuring risk in financial investments).

In your answer you **must not** use any built-in significance testing functions in R (such as `var.test`)

### Part a (2 marks)

Calculate the sample variance of the Apple log-returns and the sample variance of the SecretFund log-returns.

### Part b (4 marks)

By calculating an appropriate statistic determine if the SecretFund and Apple log-returns are statistically significantly different at the 95% confidence level. Show and explain your calculations (referring to the theorems used), and interpret your results.

**Part c (2 marks)**

Determine a confidence interval for the ratio of Apple log-returns to SecretFund log-returns. Interpret your results.

## Question 2: Linear modelling (Total 13 marks)

The following dataset describes the revenue (in £) for 30 (fictional) tech companies together with the measures of

- `advertising` - the spend (in £) on advertising
- `CEO` - the spend (in £) on the CEO's total remuneration
- `medianhourly` - the median hourly salary (in £) of the employees
- `RnD` - the spend (in £) on Research and Development
- `years_in_operation` - the number of complete financial years the company has been incorporated for.

Data for the first 3 companies is as follows:

```
##     revenue advertising      CEO medianhourly       RnD years_in_operation
## 1  587074.2     36761.5 78370.06     27.72902  23502.75                  4
## 2  996468.0    347683.8 87847.65     20.06679 107988.11                  6
## 3 1181690.7    190507.5 93124.72     27.54115 102824.69                  5
```

This data is available in the file M001 Q2 data.csv

In this question you will use a linear model to model the `revenue` in terms of the other variables. The hope is that this model could be used to predict the revenue for other companies.

### Part a (5 marks)

Write appropriate R commands to construct a least-squares linear model for `revenue` with predictors given by the other variables.

Write down an equation that describes your model (you may round the coefficients to 2 decimal places).

Which predictors are statistically significant at the 95% confidence level?

What proportion of the sum-of-squares is explained by your model?

### Part b (2 marks)

Consider a new company with the following measures:

```
##   advertising      CEO medianhourly      RnD years_in_operation
## 1      602316 72166.01     18.11379 110868.4                  0
```

This data is available in the file M001 Q2 new company data.csv

What revenue does your model predict for this new company?

### Part c (3 marks)

Re-run the linear model using only the predictors that were found to be statistically significant in part a).

What proportion of the sum-of-squares is explained by this simpler model?

Compare this to the sum-of-squares calculation you did in part a). A junior analyst claims

> "The first, more complicated, model is better because it explains a higher proportion of the sum-of-squares.' '

Do you agree? Justify your position writing a short statement either in support of the junior analyst or refuting him.

**Part d (3 marks)**

What assumptions are made when constructing a linear model?

Plot appropriate graphs to check these assumptions for your model in part c) and interpret the graphs, and make a judgement about whether the assumptions are satisfied.

## Question 3 Estimators (Total 10 marks)

Consider the one-parameter family of PDFs

$$f(x;\lambda) = \begin{cases} (\lambda+1)\,8^{-\lambda-1}x^{\lambda} & 0 < x < 8 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where the parameter $\lambda > 0$ is unknown.

Suppose the random variable $X$ has PDF $f(x;\lambda)$.

Let $X_1, \ldots, X_n$ be $n$ independent and identically distributed random variables with PDF $f(x;\lambda)$.

### Part a (2 marks)

Derive the likelihood function for $\lambda$ in terms of the random variables $X_1, \ldots, X_n$.

### Part b (2 marks)

Write R code to plot the likelihood function for the following 5 observations:

```
## [1] 6.84 7.63 6.50 6.26 2.14
```

This data is also available in the file M001 Q3 observations.csv

### Part c (6 marks)

Find the Maximum Likelihood Estimator $\widehat{\lambda}_{MLE}$ for the parameter $\lambda$ in terms of the random variables $X_1, \ldots, X_n$.

Hence find an estimate for $\lambda$ using the observations from part b).

Write a sentence explaining this estimate. What useful properties does this estimator have?

## Question 4 Combining datasets and experimental design (Total 9 marks)

In this question you will explore combining the results of two (fictional) medical trials investigating if a steroid injection could help the growth of lungs of children with asthma.

The random variable $X$ measures the difference in lung growth between participants in a control group and matched participants who were given a steroid injection.

We will assume that $X \sim N(\mu, 4)$ with an unknown value of $\mu$.

If the steroid injection **does not** affect lung growth then $\mu = 0$. We wish to run a statistical test to determine if $\mu \neq 0$.

In the first trial a study of $n = 10$ observations were made with the following results:

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 1.2$$

In the second trial a study of $n = 15$ observations were made with the following results:

$$\overline{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = 0.9$$

### Part a (2 marks)

Show that neither of these trials individually shows that $\mu \neq 0$ at the 95% confidence level.

### Part b (3 marks)

A journalist is summarising the evidence from these trials. They conclude

> "Two trials failed to find an effect of steroids on the growth of lungs. There is no statistical evidence that steroid injections work."

Is this journalist correct? Write a short paragraph in response to the journalist's claims.

Combine the data from these two trials into one big trial (i.e. with $n = 25$). Calculate the sample mean for all 25 observations, proving any formulas that you use. Is there an effect?

### Part c (4 marks)

You have been asked to plan the research for a project looking at the effects of steroids for lung growth in adult athletes.

Let the random variable $X$ measure the difference in lung growth between participants in a control group and matched participants who were given a course of steroids.

Assume that $X \sim N(\mu, 4)$ with an unknown value of $\mu$. We want to test if $\mu > 0$.

Only a result of $\mu = 0.5$ or higher would be relevant for enhancing athletic performance.

How many participants ($n$) do we need in the trial to get a 90% power at the 95% significance level? Justify your calculations.