# Group Project: Data Centre Hard Drive Reliability

300700 Statistical Decision Making, Autumn 2018

Due: Friday of Week 13 (1 June)

This Group Project analyses hard drive reliability data made publicly available by the operators of a data centre[1]. The raw data (22.7 GB in total) have been preprocessed; the data for this assignment are provided in the file `HardDisks.csv` together with this task sheet.

The file `HardDisks.csv` contains the following information for 125864 hard drives:

(1) a unique serial number;
(2) the model of the hard drive;
(3) the number of days the hard drive was operational;
(4) the mean operating temperature of the drive; and
(5) whether the hard drive was removed because it had failed.

A drive that is listed as not having failed either was still operational at the end of the data collection period, or it was removed for other reasons, for instance a capacity upgrade.

## 1 Proportion of early failing drives

We want to investigate the proportion of drives that fail in the first year of operation.

(a) [1 mark] Create a subset of the cases that is relevant for this analysis and print the number of cases in that subset.

(b) [1 mark] Compute a point estimate for the proportion of drives that fail in the first year of operation.

(c) [2 marks] Use bootstrapping to compute a 99% confidence interval for the proportion of drives that fail in the first year of operation.

## 2 Temperature and time to failure

For this part, we only consider the drives that failed. We want to analyse whether the mean operating temperature of the drive and the time to failure of the drive are associated.

(a) [1 mark] Compute and interpret the correlation between the mean operating temperature and the number of days until failure.

(b) [2 marks] Use randomisation to test at a significance level of 5% whether there is evidence that a higher mean operating temperature is associated to earlier failure.

(c) [1 mark] Interpret your findings, comparing the results from parts (a) and (b). Discuss, in particular, whether there is evidence that a higher operating temperature causes drives to fail earlier.

[1] https://www.backblaze.com

## 3 Three 2 TB drive models

For this part, we only consider 2 TB drives with the following model identifiers:

- "Hitachi HDS723020BLA642"
- "ST320005XXXX" (Seagate)
- "WDC WD20EFRX" (Western Digital)

(a) [2 marks] At a significance level of 1%, test whether there is evidence for a difference in the mean operating temperature between the three drive models **and** conduct a pairwise $t$-test.

Discuss your findings. Be specific about any differences between the models that can be inferred from the data.

(b) [3 marks] At a significance level of 1%, test whether there is evidence for a difference in the proportion of failed drives between the three drive models.

Discuss your findings. Be specific about any differences between the models that can be inferred from the data.

## Submission

One report is to be submitted by each group by the due date, containing the description of a solution to the tasks above, including any R code, and the results obtained.

**The report is to be written in R Markdown, using the template available on the unit's vUWS site.**

After editing the R Markdown file, *knit* it to Word (not HTML!) in R Studio (see http://rmarkdown.rstudio.com for full details on R Studio and R Markdown) and then convert the resulting `.docx` file to PDF using MS Word. Alternatively, you can *knit* to HTML and convert the resulting `.html` file to PDF using a web browser (with a suitable plugin or a virtual printer). If you have LaTeX installed (https://www.latex-project.org), you can also *knit* your R Markdown file directly to PDF.)

After checking that the PDF file is formatted correctly, submit it using the link *Group Project* in the *Group Project* tab on the unit's vUWS site. **Do not submit a file in any format other than PDF.** If you submit a file in another format, it may not be possible to mark your report at all, or you may lose marks due to bad formatting of code, plots or text.

The first page of your report must contain the declaration shown in Figure 1; you make this declaration by submitting your report. **Do not remove this declaration! A marker has the right not to mark your report if the above declaration is not included in the report.**

By including this statement, all authors of this work declare that:

- We hold a copy of this assignment if the original is lost or damaged.
- We hereby certify that no part of this assignment has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
- No part of the assignment has been written for us by any other person except where collaboration has been authorised by the unit coordinator.
- We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism; this software may retain a copy on its database for future plagiarism checking.
- We hereby certify that no part of this assignment or product has been submitted by any of us in another (previous or current) assessment, except where appropriately referenced, and with prior permission from the unit coordinator for this unit.
- We hereby certify that we have read and understand what the University considers to be academic misconduct, and that we are aware of the penalties that may be imposed for academic misconduct.

| Name | Student Number | Contribution (%) |
|---|---|---|
| | | |

Figure 1: Statement to be included on the first page of each submission.

# Marking Criteria and Standards

The Group Project will contribute a maximum of 15 marks towards your final mark.

The value of each of the tasks is indicated. Marks are awarded according to the following criteria:

- choice of correct method for sampling, bootstrapping, randomisation and / or analysis;
- correctness and clarity of R code; and
- correctness and clarity of analysis or interpretation.

In addition, 2 marks are awarded for the overall quality and presentation of the report.

Groups should discuss all aspects of the project: Working in a team on a statistical project is an assessable learning outcome of the unit, and the report will be treated as a team submission. The contributions by each team member must be indicated on the cover sheet.

Remember that the marker will only see what you have written, therefore, comment your R code and clearly explain all decisions made, as well as the analysis and your interpretation of the results. (Don't expect the marker to spend ages trying to figure out what you might have meant to say!)

The formatting of your report may affect its readability and the clarity of your explanations, and hence contributes to your mark.