MAT022 Foundations of Statistics and Data Science

Summative Assessment 2019/20

Summative assessment for the module is by means of a single **report** on your statistical analysis of data related to the **decathlon**, a combined event in athletics where an athlete's performance in ten track-and-field events is determined by a points system.

This form of assessment has been chosen because, as professional statisticians and data scientists, you will often be asked to investigate a data set and report on whether it contains anything useful or interesting. The assessment will also help you to prepare for writing your MSc dissertation in the summer.

Assessment type	*Weight*	*Max. length*	*Format*	*Deadline*
Report	100%	10 pages	PDF	**Friday 10 January 2020**

Your report will be assessed according to how well you are able to

- **analyse** the data set, **40%**
- **interpret** the results of your analysis, and **30%**
- **present** the results of your analysis and interpretation of the data set. **30%**

You are free to use any statistical software package to conduct your analysis (e.g. R or SPSS), and any word processing software to prepare your report (e.g. LaTeX or Microsoft Word).

1 The data

You are asked to write a report on data related to the **decathlon**. This is a combined event in athletics where an athlete's performance in ten track-and-field events is determined based on a points system. The winner of the competition is the athlete who has the most points after all ten events have been completed.

The basic decathlon data set records the performance of elite decathletes over the period from 1986 to 2006, and has been widely studied. The set consists of 7968 observations on 24 variables as shown in Table 1, and is available on Learning Central as a `.csv` file.

The basic data set is also included with the `GDAdata` package in `R` and can be loaded as follows.

```
> install.packages("GDAdata")
> data(Decathlon, package("GDAdata")
> summary(Decathlon)
```

Variable	*Description*
`Totalpoints`	Total points achieved over all 10 events
`DecathleteName`	Decathlete's name
`Nationality`	Decathlete's nationality
`m100`	Time for the 100 metres (secs)
`Longjump`	Distance jumped (metres)
`Shotput`	Distance putting the shot (metres)
`Highjump`	Height jumped (metres)
`m400`	Time for the 400 metres (secs)
`m110hurdles`	Time for the 110 metres hurdles (secs)
`Discus`	Distance throwing the discus (metres)
`Polevault`	Height achieved (metres)
`Javelin`	Distance throwing the javelin (metres)
`m1500`	Time for the 1500 metres (secs)
`yearEvent`	Year of performance
`P100m`	Points for performance in 100 metres
`Plj`	Points for performance in long jump
`Psp`	Points for performance in putting the shot
`Phj`	Points for performance in high jump
`P400m`	Points for performance in 400 metres
`P110h`	Points for performance in 110 metres hurdles
`Ppv`	Points for performance in pole vault
`Pdt`	Points for performance in discus
`Pjt`	Points for performance in javelin
`P1500`	Points for performance in 1500 metres

Table 1: The basic decathlon data set

To expand your analysis of the decathlon you are encouraged to find additional sources of data, making sure that the provenance of the sources are evaluated and discussed in your report. You are also encouraged to explore additional statistical methods that have not been discussed in the lectures and notebooks, making sure that you provide a brief description of these methods along with references to the relevant literature. You can nevertheless base your study entirely on the data set provided, it has plenty of scope for you to produce an excellent report.

2 The report

The ability to write clearly and concisely is an important professional competence. To encourage writing that is brief and to the point, your reports are limited to a **maximum of 10 pages**. It is often far more difficult to express yourself in 100 words than in 1000 words, especially when you have a lot to say, so please do not underestimate the challenge posed by this restriction. The modest page limit will also encourage you to be selective in the results you choose to present.

A suggested structure for your report is shown in Table 2. Note that the title page, abstract, table of contents and list of references will not contribute towards the page count.

- The **title page** should contain the title of your report, your name and student number, and the date on which your report was completed.
- The **abstract** should contain a short summary of the report and its main conclusions.
- The **table of contents** should list the number and title of each section against the number of the page on which the section begins.
- The **introduction** should consist of a few short paragraphs, describing the purpose of the

Title	1 page
Abstract	100 words
Table of contents	–
1. Introduction	1/2 page
2. Background	1/2 – 1 page
3. (*descriptive analysis*)	1 – 2 pages
4. (*inferential analysis*)	2 – 3 pages
5. (*inferential analysis*)	2 – 3 pages
6. Conclusion	1/2 page
References	–
Appendices	2 pages max.

Table 2: Report structure

report and providing a brief outline of its contents.

- The **background** chapter should include a brief review of any relevant literature, and provide a context for the work presented in the report.
- The report should contain a relatively short section on a **descriptive analysis** of the data, with a title chosen to reflect what the section contains.
- The main part of the report should consist of one or more sections on an **inferential analysis** of the data. Here you should formulate hypotheses, conduct statistical tests and discuss the results of these tests. The titles of these sections should reflect what the sections contain.
- The **conclusion** should consist of a few short paragraphs, providing a summary of the report and a brief outline of some ideas for future work.
- Any **references** should be typeset using the *Harvard* referencing style.
- The report may contain a single **appendix** for large figures and tables.

3 Assessment criteria

Detailed assessment criteria are shown in Table 3.

4 Guidelines for writing reports

The golden rule when writing is to always **think of the reader**. For scientific reports, readers will typically want to read something interesting and to learn something in the process.

What do we mean by interesting?	
Not interesting	The average exam marks of statistics and data science students.
Quite interesting	The average marks of male students, the average marks of female students, and the results of a test of whether any difference is statistically significant.
Very interesting	The average marks of male students, the average marks of female students, a statistical test of whether any difference is significant, and some speculation about why there is a significant difference, or alternatively why there is not.

Level	Analysis (40%)	Discussion (30%)	Presentation (30%)
Distinction (70–100)	Hypotheses are interesting and original. Methods are appropriate and applied carefully and precisely. An interesting descriptive analysis is included and reported correctly.	Inferences are valid and supported by evidence. Original and interesting conclusions are articulated. There is some shrewd speculation about possible causal factors.	A high standard of writing is maintained throughout. The narrative is clear, coherent, eloquent and refined. Figures and tables are used creatively.
Merit (60–69)	Hypotheses are formulated correctly. Methods are appropriate and applied correctly. A moderately interesting descriptive analysis is included and reported correctly.	Inferences are valid and supported by evidence. Interesting conclusions are articulated. There is some speculation about possible causal factors.	A good standard or writing is maintained throughout. The narrative is clear and coherent. Figures and tables are used to illustrate the narrative.
Pass (50–59)	Hypotheses are formulated correctly. Methods are applied correctly for the most part. A descriptive analysis is included and reported correctly.	Inferences are mostly valid and supported by some evidence. Some relatively interesting conclusions are articulated.	An acceptable standard of writing is maintained throughout. The narrative is lacklusture and sometimes unclear. Figures and tables do not always illustrate the narrative.
Fail (0–49)	The analysis is bland and almost entirely descriptive.	Inferences are invalid or not supported by evidence. There is little of any interest.	The report is poorly written. The narrative is disjointed and hard to follow.

Table 3: Assessment criteria

Audience. The target audience for your report is this year's cohort students on the *Foundations of Statistics and Data Science* module, so you can assume that your readers are familiar with the methods and terminology established within the lectures and notebooks. If you choose to use methods that have not been covered in lectures, you must ensure that any new terms are properly defined, and references to the relevant literature included.

Analysis. The reader shoud be satisfied that you have performed your analysis correctly, and in particular that you have verified the conditions that are necessary to apply the various methods. Your methods should be introduced with a brief summary of their main features, but technical details should not be discussed at length, although you might consider providing the interested reader with references to the relevant literature.

Navigation. Do not assume that the reader will read the report from start to finish, as one might read a novel. Reports should be made easy to navigate using numbered sections and subsections together with cross-referencing. Once you have written a first draft, it will need careful editing before it becomes a coherent and polished report. This stage always takes longer than you think!

Scientific writing. For scientific reports we aim for a style of writing that is *clear* and *concise*. Make sure that sentences are unambiguous and that a good standard of writing is maintained throughout the report.

- Sections should not start abruptly with the subject matter, but rather with an introductory sentence or short paragraph. Sections should also end with concluding sentence or short paragraph.
- *All* figures and tables must be numbered and have captions. Figures or tables that are not mentioned at least once in the text should not be included.
- A *qualified statement* is one that express some level of uncertainty about its own accuracy, and should always be used when drawing conclusions from the results of a statistical analysis, and especially when speculating about possible causal factors. Common phrases that indicate qualified statements include "*This suggests that ...*", "*It appears that ...*", "*We might conclude that ...*", "*There is some evidence to indicate ...*" and so on.
- Be careful with *florid turns of phrase*, whatever their merit as literature. Academic reports are primarily a way of communicating information, and care must be taken to accommodate readers from diverse backgrounds, including non-native English speakers. Reports should use everyday words and simple grammatical structures as far as possible.

Plagiarism

The basic decathlon data set has been widely studied and you will find plenty of material online about this data. Plagiarism is presenting other people's work and ideas or ideas as your own, by incorporating it into your work without full acknowledgement. The need to acknowledge others' work applies not only to text, but also to computer code, figures, tables etc. You must also attribute text, data, or other resources downloaded from websites. Following submission your report will be analysed by the *TurnitIn* software, and any report in which plagiarism is detected will receive a mark of zero.

Please submit your report via Learning Central on or before **Friday 10 January 2020**.