Department of Statistical Science

STAT0017 IN-COURSE ASSESSMENT 1 2018-2019

Deadline: 5pm Friday 22nd March 2019

Your report must be your own work. It will be marked anonymously. **Do not put your name on your report.**

Rules for submitting your report

- You will submit, via a link on the Topic 1 – ICA section of the STAT0017 Moodle page, **one PDF file.** Further guidance is given later in these instructions.
- By ticking the submission declaration box in Moodle you are agreeing to the following declaration.

  Declaration: I am aware of the UCL Statistical Science Department's regulations on plagiarism for assessed coursework. I have read the guidelines in the student handbook and understand what constitutes plagiarism. I hereby affirm that the work I am submitting for this in-course assessment is entirely our own.

- The Turn-It-In® plagiarism detection system may be used to scan your submission for evidence of plagiarism and collusion.
- Any plagiarism will normally result in zero marks for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in Departmental Student Handbooks. The relevant excerpt from the Statistical Science handbook appears at the start of these instructions and is also posted in the Topic 1 – ICA section of the Moodle page.
- Late submission will incur a penalty unless there are extenuating circumstances (e.g. medical) supported by appropriate documentation. Penalties are set out in the latest editions of the Statistical Science Department student handbooks, available from the departmental web pages.
- Failure to submit this in-course assessment may mean that your overall examination mark is recorded as “non-complete”, i.e. you will not obtain a pass for the course.
- You will receive, via Moodle, feedback on your work and a provisional grade – grades are provisional until confirmed by the Statistics Examiners' Meeting in June 2019.

Paul Northrop, p.northrop@ucl.ac.uk, 7/2/2019

PLAGIARISM AND COLLUSION – EXTRACT FROM DEPARTMENTAL STUDENT HANDBOOKS

Plagiarism means attempting to pass off someone else's work as your own, while collusion means passing off joint work as your own unaided effort. Both are unacceptable, particularly in material submitted for examination purposes including exercises done in your own time for in-course assessment. Plagiarism and collusion are regarded by the College as examination irregularities (i.e. cheating) and are taken extremely seriously. UCL uses a sophisticated detection system (Turnitin®) to scan work for evidence of plagiarism and collusion, and the Department reserves the right to use this for assessed coursework. This system gives access to billions of sources worldwide, including websites and journals, as well as other work submitted to the Department, UCL and other universities. It is therefore able to detect similarities between scripts that indicate unacceptable levels of collusion, as well as material taken from other sources without attribution.

If plagiarism or collusion are suspected, on the basis either of the Turnitin® software or other evidence, it can be dealt with informally only in the case of first offences committed by first year students. All other cases must be dealt with formally, which involves adjudication by a departmental panel and/or College Examinations Irregularities panel. If the panel finds that an offence of plagiarism or collusion has been committed, a penalty will be imposed. Penalties depend on the severity of the offence, and range from being awarded zero marks for the work in question up to exclusion from all further examinations. They can also include a formal reprimand, which will be entered on the student's departmental and College records.

What isn't acceptable?

Students sometimes find it difficult to know what counts as plagiarism or collusion. The following list is not exhaustive, but gives some indication of what to avoid. It is based on guidelines developed by Nick Hayes of the UCL Pharmacology Department. You may **NOT**:

- Create a piece of work by cutting and pasting material from other sources (including websites, books, lecture notes and other students' work).
- Use someone else's work as your own. This includes, but is not limited to:
    - Making notes while discussing an assessment with a friend, and subsequently using these as the basis for all or part of your submission.
    - Telephoning another student to discuss how best to carry out a particular piece of analysis.
    - Employing a professional ghostwriting firm or anyone else to produce work for you.
- Use somebody else's ideas in your work without citing them.
- Ask a lecturer in the department for help with assessed work, unless you make it clear to them that the work is assessed.
- Help another student with their assessed work. If you do this, you will be deemed to be guilty of an examination irregularity.

What is acceptable?

The following practices do not constitute plagiarism / collusion:

- Quoting from other people's work, with the source (e.g. book, lecture notes, website) clearly identified and the quotation enclosed in quotation marks.
- Summarising or paraphrasing other people's work, providing they are acknowledged as the source of the ideas (again, usually this will be via a reference to the book, journal or website from which the information was obtained).
- Asking the course lecturer for help with difficult material, providing it is clear that the question is in connection with the assessment. The lecturer will be able to judge for him or herself what is an appropriate level of assistance.

Some examples

Unfortunately, each year there are some students in the Department of Statistical Science who submit work that contravenes the regulations. The consequences can be severe.

Example 1: Final-year student A had a lot of coursework deadlines in the same week as an important job interview. One of the coursework deadlines was for an extended piece of data analysis, set two weeks previously. Because of his other commitments, student A did not start this piece of coursework until shortly before the deadline, at which point he discovered that he did not have enough time to do it. He asked student B for help. The result was that both students submitted essentially identical work using exactly the same computer output. A departmental panel was convened to investigate the matter. The panel suggested that student B had passed electronic material (computer output and graphics files) to student A, who had pasted this material straight into his own submission. Although student A admitted asking student B for help, both students denied exchanging electronic material. They were, however, unable to explain how the same electronic files came to appear in both submissions. As a result, the allegation was upheld and both students were penalised. Student A was recorded as "non-complete" for the course in question (this meant that he had no possibility of passing it that year), and student B was given a mark of zero for the coursework component.

Example 2: Students C and D both had to submit some computer code for an assessment, which was worth one third of the total mark for a course. There was considerable flexibility in how to go about the assessment. Although the students submitted code that looked very different, closer inspection revealed that they were carrying out the same procedures in more or less the same order, and that the methods they used to carry out these procedures were essentially the same. Further, these procedures and methods were not used by other students in the class. On investigation, it transpired that the students had discussed the assessment over the phone while sitting in front of their computers. This is unacceptable, and as a result the marks of both students for this piece of assessment were halved.

Example 3: The in-course assessment for a particular module was organised as a multiple choice exam taken via Moodle outside of lessons. Each student could attempt the one-hour exam at any time of their choosing within a ten day window, but were clearly advised that they must work alone. After the exams had been graded, it was noticed that students E and F had given identical answers to every question (including incorrect answers). Inspection of the Moodle logs revealed that the students had started and finished their attempts at exactly the same time, using IP addresses that were traced to adjacent PCs in the same computer cluster. Students E and F admitted colluding on the in-course assessment and were both given a mark of zero.

How to avoid plagiarism and collusion

If you are found to have committed an offence of plagiarism or collusion, it makes no difference whether or not you intended to do so. Ignorance is no excuse. To avoid committing an offence, a useful rule of thumb is: if in doubt, don't do it. Make sure that any work you submit is your own unaided effort. More specific guidance is as follows:

- Plan your work schedule carefully, to allow enough time to complete each piece of assessment.
- If you have genuine problems in meeting a deadline, don't take the easy way out and borrow a friend's work. Discuss your difficulty with the course lecturer in the first instance.
- If you are stuck with an assessment, don't ask another student for help. Discuss it with the course lecturer.
- If another student asks you for help with an assessment, or asks to see your work, suggest that they approach the course lecturer instead. Remember: if somebody else copies or uses your work, you will be penalised as well, even if you didn't expect them to use your work in this way.

STAT0017 ICA 1 2018-2019

This assessment contains two separate parts: Part A and Part B.

- Part A is Section 1 below. It is based on a mathematical exercise relating to the extremal types theorem.
- Part B is Sections 2–4 below. It requires you to use R to perform extreme value modelling of data and to comment on these analyses and associated issues.

A template R markdown file is provided in the Topic 1 - ICA section of the Moodle page. This file is set up to produce (using Rstudio's Knit button) a PDF with a structure that should make it easier for you to prepare your ICA submission. Download this file and rename it to `SNxxxxxxxx.Rmd`, where `xxxxxxxx` is your student number. Also add your student number to the author field.

You must submit **ONE PDF file** that contains your submission for boths Parts A and B. You may do this in any way that you find convenient. For example, you could produce a PDF file for Part A (a scan of handwriting and/or a LaTeX or Word document) and combine it with the PDF file that you produce, using R markdown, for Part B. If you are using LaTeX then you could consider including your answers to Part A in the template R markdown document, to create one PDF file from the outset. I have included a Section in the R markdown template in case you wish to do this.

Your PDF file must contain no more than 25 pages. Anything beyond 25 pages will not be marked. This is a generous limit, which you should not aim to reach: taking up more space will not gain more marks.

The numbers in square brackets indicate the relative weight attached to each part of the assessment.

Part A: Extremal Types Theorem

1 A distribution in the Gumbel domain of attraction

Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with distribution function

$$F(x) = P(X \leqslant x) \;=\; \begin{cases} 1 - \mathrm{e}^{1/x} & \text{for } x < 0, \\ 1 & \text{for } x \geqslant 0. \end{cases} \tag{1}$$

Let $M_n = \max(X_1, \ldots, X_n)$.

(a) Use the von Mises' condition, on slide 37 of part 1 of the lecture slides, to show that this distribution function is in the domain of attraction of the Gumbel family. Also use the theory on slide 37 to find a_n and b_n for which $(M_n - b_n)/a_n$ achieves this limit as $n \to \infty$. [6]

(b) Suppose that one of your colleagues taking STAT0017 says: "I am surprised that a distribution with a finite upper end point of 0 is in the domain of attraction of the Gumbel family, which has an upper end point of infinity." What would you say to convince them that this result is correct and explain to them why this occurs in this example? [7]

(c) Suppose that we have a random sample $m_1, \ldots, m_N$ from the distribution of M_n. Explain why we would fit a GEV to these data even if we believed strongly that the variables contributing to these maxima were sampled from a distribution with distribution function (1). [6]

(d) Suppose that $n = 365$ and that we are in the fortunate position of having an extremely large sample of maxima, because $N = 10,000$. If we fit a GEV distribution to these data, using, for example, maximum likelihood estimation, what do we expect the estimate of the shape parameter ξ to be, approximately? [6]

Part B : extreme value modelling of significant wave height data

You need to imagine that you are a statistical consultant to an oil company. The company is planning to build a new off-shore oil platform at a secret location in the northern North Sea. They need to make an assessment of the meteorological and oceanographic ('metocean') conditions to which the oil platform may be subjected during the time for which is planned to be in operation. This enables them to decide how much money they need to invest in strengthening the structure of the oil platform.

To assist with this the company has collected data from their location of interest relating to metocean variables. One of these variables is significant wave height, denoted by H_s. Significant wave height is a measure of sea surface roughness: the larger is H_s the rougher the sea. It is a key variable used by the company in judging whether an oil platform can operate safely. To make this judgement it is necessary to consider how large the largest value H_s experienced by the oil platform might be.

The company has also collated data on two climate indices: the Arctic Oscillation (AO) and the North Atlantic Oscillation (NAO) (see also this discussion of the NAO). They have provided these data because historically they exhibit a statistical association with weather patterns in the Northern Hemisphere. The raw NAO and AO index values (from which the summaries in the dataframe `wm` detailed below are calculated) are the daily values provided by the National Oceanic and Atmospheric Administration (NOAA) service in the US, downloaded from here (NAO) and here (AO).

Sections 2–4 below are based on analyses of two R dataframes: `wm` and `pot`, which are available from the Topic 1 - ICA section of the STAT0017 Moodle page. In each of these dataframes the values of H_s are based on time series of the roughness of the sea over disjoint 3-hour periods. Some pre-processing has been performed by the company.

- They have used an in-house algorithm to identify separate storms from the data and have retained only the largest H_s value (the *storm maximum*) observed during each storm. In short, they have *declustered* the data so that it is reasonable to treat the storm maxima as being independent.
- They have retained in the data only observations observed during winter, where, for their purposes, winter is defined as 1st October - 31st March inclusive. They have done this because in their location of interest the largest values of H_s occur almost exclusively in winter.

The dataframe `wm` ('winter maxima') with 56 rows and 6 variables:

- `Hs`: the largest storm maximum H_s value (in metres) observed over a given winter;
- `waterYear`: the year in which this winter ends (a *water year* starts on 1st October);
- `NAOmean`: the mean value of the NAO index over a winter;
- `NAOmax`: the maximum value of the NAO index over a winter;
- `AOmean`: the mean value of the AO index over a winter;
- `AOmax`: the maximum value of the AO index over a winter;

The dataframe `pot` ('peaks-over-threshold') with 595 rows and 7 variables:

- `Hs`: the storm maximum significant wave height (in metres) observed over a given storm;
- `year`: the year in which a storm maximum was observed;

- month: the month in which a storm maximum was observed;
- day: the day on which a storm maximum was observed;
- date: the date (yyyy-mm-dd format) on which a storm maximum was observed;
- dayOfYear: the day on the year on which a storm maximum was observed;
- sdd: the *seasonal degree day*, the day on which a storm maximum was observed, for a standardized year of 360 days.

In both wm and pot the time period covered by the data is 1/10/1954 to 31/3/2010, that is, the winters from 56 water years.

Your tasks

Your client, the oil company, is focused mainly on the quantity H_s^{100}, the largest storm maximum H_s value to be observed at their location of interest over the next 100 years. In particular they require:

A. an interval estimate of H_s^{100} with a coverage probability of 95%.

B. an estimate of the value that is exceeded by H_s^{100} with a probability of 1%.

C. a plot of the c.d.f. of H_s^{100}.

You will be judged on how you tackle the specific tasks that are detailed below, on the clarity and correctness of the explanations and comments that you make in relation to these tasks and on the judgement that you show in deciding what to include in your report. However, you should keep your client's requirements in your mind as you perform your analyses.

The template R markdown file (*remember to rename it!*) is set up with sections that relate to Sections 2 to 4 below. You need to add R code to execute and (in the comments sections) comments that explain what you learn from the output. The R code from the computer practicals will be useful, but you may also need to use R's help system to find out more about the functions that you used.

There is no need to explain in detail what your R code does or how it works. However, if you would like to include some **very brief** comments to help you keep track of what you are doing then you are welcome to do so.

You may find that there is a lot of white space surrounding the plots in your PDF file. Don't worry about this. However, you should try not include in the PDF file that you submit a lot of unnecessary output. For example, if a function that you call outputs a large object then you should avoid the whole object appearing in your output: include only the parts that you need. You may find useful the show = FALSE argument to the gev.fit() function in the ismev package.

2 Exploratory analysis

Produce, and comment on, simple graphs to explore the behaviour of H_s in the wm and pot datasets and, for wm only, and its relationship with the summaries of NAO and AO. [10]

3 Extreme value (EV) modelling of H_s

Perform the following extreme value analyses of the H_s data, using both maximum likelihood and Bayesian inferences in each case.

3.1 in `wm`, using a GEV model for winter maxima, and

3.2 in `pot`, using a GP model for threshold excesses (and a binomial model for the number of threshold exceedances).

[*If you use the* `revdbayes` *package to perform the Bayesian analysis in 2. then you may see a convergence warning message. You can ignore this: the convergence is fine.*] [25]

3.3 Reporting to your client

Choose one of the (four) analyses that you have performed, and use it to provide information to satisfy your client's 3 requirements: A, B and C. Explain why you chose this analysis. [*There is no absolutely correct choice: marks will be awarded based on your explanation, not for the choice itself.*] [15]

4 EV regression modelling of winter maximum H_s on NAO

4.1 Build a GEV regression model

Use the `wm` dataset to build a GEV extreme value regression model with winter maxima of H_s as the response. The potential covariates are the other variables in `wm`.

Use only maximum likelihood estimation. This is because (a) the `evdbayes` package can perform Bayesian inferences only for a certain type of GEV model (with covariate effects only in the location parameter μ), and (b) you should be familar with the general approach to building a model when fitting using maximum likelihood estimation.

Include in your submitted PDF file only information about the final model that you select and a brief description of how you selected this model. [15]

4.2 Inference for H_s^{100}

Your client had been hoping to provide you with projections of the NOA and AO indices for the next 100 years, based on a climate model that they had been developing, in order that you could use these to make inferences about H_s^{100}. However, they have a problem with their model and were not able to do this.

Suppose that you do have values of the winter summaries of NAO and AO in `wm` projected for each of the next 100 years. Explain how you would use these data, and the model that you selected in Section 4.1, to estimate the probability $P(H_s^{100} \leqslant x)$, for some value x of interest.

You should use some mathematical notation, which you should define clearly, and give enough information that it someone wished to write computer code to estimate $P(H_s^{100} \leqslant x)$ then they could use your explanation to guide them. [10]