# G3/M3 Assessment 2 — Instructions

1. Answer **both** questions.

2. For Question 1 you should hand in a paper copy of your report for part (d), and you should also upload an electronic copy of your report to the course Moodle page. **You are NOT required to submit your R script for this question**.

3. For Question 2 you should submit:

   - an electronic copy of your `Surname.r` file containing your R script (see below),
   - a printout of your `Surname.r` file containing your R script (see below),
   - a printout of the graph in part (b) that is produced by your script (black and white copy is sufficient), and
   - a printout of your `Surname_out.txt` file (see below) containing output to parts (d), (e) and (f) produced by your R script.

   Your R script for this question should be saved in a file named `Surname.r`, where `Surname` is your surname. For example, if your name is Karl Pearson, your R script should be saved in the file `Pearson.r`.

   Your script should create an output file called `Surname_out.txt`, where again `Surname` is your surname. This file should contain the output from parts (d), (e) and (f) and it should include text and comments indicating what the results are (produced via appropriate use of the `cat()` function in your code). Any output should correspond exactly to what appears on the screen when sourcing your script file.

   Your program should be **well commented**. It should consist of a header section summarising the logical structure, followed by the main body of the script. The main body should itself contain comments. You should clearly indicate the question and part numbers in both your code and output.

4. Paper copies of your answers and printouts should be handed in to the Statistical Science departmental office. Your answers to **both** questions should be submitted in a **single stapled document** (not in loose pages) that is clearly identified with your name. **Remember to complete a plagiarism declaration, and to attach it to your work**.

5. Electronic copies of your report for Question 1, and your script for Question 2, should be submitted via the Moodle page for the course — use the link "ICA2: Click here to submit your assignment".

# 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.

# G3/M3 Assessment 2 — Hints

1. In general, there is not a single 'right' answer to each question. To obtain a good mark you should approach the questions sensibly and justify what you're doing. Credit will be given for code that is clear and readable, while code that is inadequately commented will be penalised. You might like to use scripts `cosapprox.r` (Lab 1) and `tablet.r` (Lab 3) as models.

2. Question 1 is designed to test your ability to use the computer to learn about a real data set. This will be assessed not only on your computing skills, but also on your ability to carry out a sensible and informed statistical analysis: material from your other courses (in particular STATG001) will be relevant here. To earn high marks for this question, you need to take a structured and critical approach to the analysis and to demonstrate appropriate judgement in your choice of material to present.

3. In Question 2, make sure that the output for parts (d)–(f) is labelled appropriately so that the individual analyses can be identified. See question 2 at the end of Lab 3 for an example of how to do this.

4. Do not edit your `Surname_out.txt` file in any way before printing it for submission. Marks will be deducted if your printout does not correspond *exactly* to the results we obtain when we run the electronic versions of your scripts.

5. More credit will usually be given for code that is more generally applicable, rather than tailored to a particular situation or set of data. For example, if you were asked to print out the mean age of a group of people, you could do either of the following:

   - Calculate the mean before you write your final script, and then insert a line

     ```
     cat("Mean age is 25.3\n")
     ```

     (or whatever the mean happens to be) into your script.
   - In your script, create an object (say `xbar`) that holds the mean age, and then insert the line

     ```
     cat(paste("Mean age is",xbar,"\n"))
     ```

     into your script.

   The second approach is clearly more general and will earn more credit, since it will work for other similar data also.

6. All graphs should be clearly and appropriately labelled (giving units of quantitative variables), titled and formatted. By 'appropriately formatted' we mean, for example, that axis scales should be well chosen.

7. Both questions carry equal marks.

8. Refer to the feedback you received on in-course assessment 1.

# G3/M3 Assessment 2 – Marking guidelines

Questions 1 and 2 are each marked out of 30.

The marks for Question 1 are **roughly** subdivided into the following components.

1. Exploratory analysis (10 marks): investigation and commentary of initial statistical properties, relationships, and anything of note which helps justify your choice of graphs and modelling strategy.

2. Graphical presentation (5 marks): appropriate choice of graphs and formatting.

3. Modelling strategy (10 marks): marks here will be based on a structured, justified, well-principled approach with clear and concise discussion.

4. Interpretation of final model (5 marks): commentary on how good the model is and what it means in reality in the context of the third part of Question 1d.

The marks for Question 2 are **roughly** allocated as follows:

1. File handling and plotting, Parts (a) and (b) (5 marks): read in file; calculate and print quantities; be able to produce and format graph according to instructions.

2. Negative log-likelihood function, Parts (c) and (d) (10 marks): write `negbinll()` function that works and follows good programming practice (is usable, extensible, etc); has appropriate inputs and outputs; carry out appropriate testing.

3. Optimisation, Parts (e) and (f) (10 marks): considerate and correct use of `nlm()` function, correctly compute standard errors.

4. Style (5 marks): efficient, elegant, extensible, well-laid out, readable code. See also examples `cosapprox.r` (Lab 1) and `tablet.r` (Lab 3) for inspiration.

Marks will be deducted for code that a user would find 'difficult' to use. The 'user' here can either be (i) someone who cannot code and only knows how to run an R script and expects something meaningful to be produced on their screen or written to file; (ii) a fellow developer who would like to not only run your code but also understand how it works with a view to maybe building some of their own code on top of it. Generally, both of these user types should find your code useful and easy to use in order for you to get good marks.

# STATG003/M003 Assessment 2 — Questions

**1.** The file `vitd.dat` contains data from 30 people. Vitamin D can be manufactured by human skin being exposed to sunlight or be obtained from food sources. Often the overall health of a person, such as obesity and fitness can also be reflected in nutrition levels. There are four quantitative variables: the vitamin D status (measured in ng/ml serum concentration and denoted by `vitd`), the body mass index (BMI) of the person (denoted by `bmi`), average time spent outside in hours per week (denoted by `outside`) and average time spent exercising in hours per week (denoted by `exercise`). In addition there is an indicator variable `suppl` denoting whether the person takes a vitamin D supplement (1=yes, 0=no). Epidemiologists are interested in how a person's vitamin D status depends on the BMI, time spent outside, exercise and supplements.

(a) Download the file `vitd.dat` from the G3/M3 Moodle page. Read the data into R using `read.table` with the argument `header=TRUE`.

(b) Obtain summary statistics for each quantitative variable and make useful plots of the data — i.e., that are relevant to the objectives of the study. Such plots may include, but are not necessarily restricted to, pairwise scatter plots with different plotting symbols for those who do or do not take supplements. Put plots together in a single figure where appropriate and consider the possibility of using log scales for the quantitative variables.

(c) Find a linear model that enables `vitd` to be predicted from the other variables and that is not more complicated than necessary. You may wish to consider using log transformations of one or more of the explanatory variables or of the response variable. You should consider a wide enough range of models to make your choice of model convincing and use appropriate diagnostics to assess them. But ultimately you are required to recommend a single model that is suitable for interpretation and to justify your recommendation.

(d) Write a brief report on your analysis in three sections:

I Describe briefly what you found in your exploratory analysis in part (b)

II Describe briefly (without too many technical details) what models you considered in part (c) and why you chose the model you did, and

III State your final model clearly and describe it in words. Remember to include an estimate of the error standard deviation and say what this means also.

Your report should not include all of your R commands and output, but it should include *some* R commands and output (for example, relating to your final choice of model) and your most useful graphs. It should be limited to at most three pages of text (including any output) and two pages of graphs. Your report should be at a level that can be understood easily by somebody with an MSc in Statistics.

**2.** The file `cells.dat` contains estimates of the number of cells contained in different compartments. The cell counts $X$ (denoted by `ce` in the file) are assumed to follow a Negative Binomal distribution with parameters $\tau$ and $p = \frac{\tau}{(\tau+\mu)}$, say, with probability mass function

$$P(X = x) = \binom{(\tau + x - 1)}{x} p^{\tau}(1-p)^{x}$$

for $x = 0, 1, 2, \ldots$. The mean of this distribution is $\mu$, and the variance is $\frac{\tau(1-p)}{p^2}$.

For each value $x$ in column `ce`, the corresponding value in column `comp` is the number $f_x$ of compartments containing $x$ cells. It is required to estimate $\tau$ and $\mu$ from these data by maximum likelihood. The log likelihood function is

$$l(\tau, \mu) = \sum_{x=0}^{\infty} \Big( f_x \log P(X = x) \Big),$$

which is defined for $\tau \geq 0$ and $\mu \geq 0$.

(a) Download the data `cells.dat` from the G3/M3 Moodle page. Read it into R using `read.table` with the argument `header=TRUE`.

(b) Obtain and print out the total number of compartments, along with the mean number of cells per compartment. Plot a barchart showing the distribution of cell counts, with the mean printed on the barchart. Label your figure and axes informatively.

(c) Write a function called `negbinll` that takes two arguments

(i) `params`, a vector containing the values of the two parameters $(\tau, \mu)$, and

(ii) `dat`, a matrix of the data pairs,

and returns the negative log-likelihood, $-l(\tau, \mu)$. (Hint `R` function `dnbinom` maybe useful in computing the negative log-likelihood.)

(d) Use your function `negbinll` to evaluate and print out the negative log-likelihood for the data in `cells.dat` for a few sensible values of $\tau$ and $\mu$.

(e) Use the R function `nlm` to find and print out the maximum likelihood estimates of $\tau$ and $\mu$ for the data in `cells.dat` by minimising the negative log likelihood.

(f) Obtain and print out approximate standard errors for these estimates.