Ve401 Probabilistic Methods in Engineering

Summer 2020 Term Project 1

Date Due: 11:00 PM, Friday, the 3rd of June 2020

Group Work

You will be divided into groups of *3–4 students* each.

> Each group member must be familiar with and have contributed to each part of the project report. **You may not divide up the work in such a way that only certain members are involved with certain parts.** In the event of an Honor Code violation (plagiarism or other), all members of the group will be held equally responsible for the violation. Exceptions may only be made, at my discretion, in exceptional situations.

It is therefore all group members' duty to ensure that all collaborators' contributions are plausibly their own and to check on all collaborators' work progress and verify their contributions within reason.

Project Report

The term project will be submitted **electronically only** as a typed report of **10-15 pages** in length. Handwritten submission will not be accepted! It is recommended that you use a professional type-setting program (such as LaTeX) for your report. Unless you are able to ensure a unified font size and style for formulas and text in Microsoft Word, use of Word is *not recommended*.

Your report should have the appearance, style and contents of a professional report. It should be comprehensible without reference to this document and should be comprehensible by any other student in this course. It is strongly suggested that all members of the project team proof-read the report before submission. **The report should not look like the solution to an assignment.** Do not structure the section titles as "Answer to Question i)" or similarly.

Grading Policy

This term project accounts for 15% of the course grade; it will be scored based on

- **Form (3 points):** Does the report contain essential elements, such as a cover page (with title, date, list of authors), a synopsis (abstract giving the main conclusions of the project), table of contents, clear section headings, introduction, clear division into sections and appendices with informative titles and bibliography (if applicable)? Are the pages numbered? Are the text and formulas composed in a unified font? Are all figures (graphs and images) clearly labeled with identifiable source?
- **Language (3 points):** Is the style of english appropriate for a technical report? Do not treat the project as an assignment and simply number your results like part-exercises. Your text should be a single, coherent whole. The text should be a pleasant read for anyone wanting to find out about the subject matter.

  Errors in grammar and orthography (use a spell-checker!) will be penalized. Make sure that the report is interesting to read. Avoid simply repeating sentences by cut-and-paste.
- **Content (9 points):** Are the mathematical and statistical methods and deductions clearly exhibited and easy to follow? Are the conclusions well-supported by the mathematical analysis? It is important to not just copy calculations from elsewhere, but to fully make them your own, adding details and comments where necessary.

All group members will generally receive the same grade for the term project. Exceptions are possible in certain circumstances, such as a group member not contributing to the project.

On Plagiarism

Study JI's Honor Code carefully. **Any** information from third parties (books, web sites, even conversations) that you use in your project must be accounted for in the bibliography, with a reference in the text. Follow the rules regarding the correct attribution of sources that you have learned in your English course (e.g., Vy100, Vy200). All members of a group are jointly responsible for the correct attribution of all sources in all parts of the project essay, i.e., any plagiarism will be considered a violation of the Honor Code by all group members. Every group member has a duty to confirm the origin of any part of the text.

The following list includes some specific examples of plagiarism:

- Use of any passage of three words or longer from another source without proper attribution. Use of any phrase of three words or more must be enclosed in quotation marks ("example, example, example"). This excludes set phrases (e.g., "and so on", "it follows that") and very precise technical terminology (e.g., "without loss of generality", "reject the null hypothesis") that cannot be paraphrased,
- Use of material from an uncredited source, making very minor changes (like word order or verb tense) to avoid the three-word rule.
- Inclusion of facts, data, ideas or theories originally thought of by someone else, without giving that person (organization, etc.) credit.
- Paraphrasing of ideas or theories without crediting the original thinker.
- Use of images, computer code and other tools and media without appropriate credit to their creator and in accordance with relevant copyright laws.

Benford's Distribution

This project concerns itself with the distribution of digits in "real life" numbers. In 1881, Simon Newcomb [1] published a remark observing that "natural numbers" (those actually occurring, e.g., in physical constants or in dailty life) appear to have a non-uniform distribution of digits. At first, one might think that the initial digits of numbers occur with equal frequency, i.e., 1/9 of all numbers encountered begin with a 1, 1/9 begin with a 2, 1/9 begin with a 3, etc. While it is clear that a discrete uniform distribution can not apply if numbers are ohysically constrained (for example, the height of any person measured in cm will most often begin with the digit "1") this effect is even observable if no such constraint exists.

Frank Benford independently noticed this effect in a book of logarithm tables (used for calculations) where the initial pages were much more worn by use than the later pages. He was the first to systematically investigate the effect in 1938 [2]. The observed distribution of digits is now known as *Benford's law* or *Benford's distribution*.

There are several different approaches to Benford's law and the law can in fact be formulated in different settings. Here, we focus on the occurrence of numbers in real life. One basic argument is the following:

> Given a collection of naturally occurring numbers whose size is not constrained by outside effects, the distribution of the leading digits should not depend on the units of measurement used.

For example, if the numbers are length measurements, the proportion of 1s, 2s, 3s, etc. should be the same, whether the lengths are measured in meters, in feet or in any other unit system. Of course, individual lengths will have different numerical expressions, but the overall distribution of leading digits should not be affected by unit choice. This is a *scaling argument*, since it claims that the distribution of digits should be invariant under re-scaling, i.e., a change of units of measurement.

Your project should introduce the Benford distribution and cover the following items:

i) Show that if the leading digits of a discrete random variable follow a discrete uniform distribution (each digit occurs with probability 1/9) then this distribution is not independent of re-scaling.

ii) Take a table of physical data that spans several orders of magnitude and is imbued with units. Create a histogram of the frequencies of the numbers 1-9 occurring as leading digits. Then change the physical units in a non-trivial manner (i.e., not simply by a factor of powers of 10) and re-draw the histogram using the numerical values in the new units.

(An example of such data is the table of values of the shear modulus for the solid elements in the Elastic Properties of Elements [3]; however, you should find your own suitable data!)

Compare your histograms to Benford's distribution (purely qualitatively). Comment on the result.

iii) One argument as to *why* Benford's law should hold is that the occurrence of digits should follow a distribution that does not change when units are changed and the data is rescaled. Pinkham [4] gave an argument that purported to show that this scale invariance implies Benford's law. However, Pinkham's argument had a serious flaw. Theodore Hill [5] gave the first mathematically rigorous proof in 1995 and discusses the problems that previous attempts at proofs had. Summarize Pinkham's argument and discuss why it is flawed.

iv) An extensive survey article on Benford's distributionw as published by Berger and Hill [6] in 2011. Several approaches are summarized there, including scaling invariance, base invariance and a "law of large numbers"-like approach. Choose one of these approaches and work through the arguments in the proof that chows that Benford's distribution must occur in the corresponding circumstances.

You should introduce and explain all necessary background information (e.g., "σ-algebra", "significand", etc.) and write down the statement and the proof in language that any sophomore or junior student at JI could understand, given the knowledge you currently have learned in Ve401.

To re-iterate: in the end, the argument should be crystal clear, with every small step and conclusion adequately explained, every term used should be defined or well-known and the proof as a whole should be easy to read and understand by a sophomore student.

One of the applications of Benford's law is fraud detection - when numbers are invented by people (e.g., to cover up fraud in accounting) these numbers rarely follow Benford's law. An accessible overview and example can be found in [7].

References

[1] Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1):39–40, 1881. `http://www.jstor.org/stable/2369148`.

[2] Frank Benford. The law of anomalous numbers. *Proc. Amer. Philosophical Soc.*, 78:551–572, 1938. `http://www.jstor.org/stable/984802`.

[3] Wikipedia. Elastic properties of the elements (data page) — wikipedia, the free encyclopedia, 2013. `https://en.wikipedia.org/w/index.php?title=Elastic_properties_of_the_elements_(data_page)` Web. Accessed June 29th, 2015.

[4] Roger S. Pinkham. On the distribution of first significant digits. *Ann. Math. Statist.*, 32(4):1223–1230, 12 1961. `http://dx.doi.org/10.1214/aoms/1177704862`.

[5] Theodore P. Hill. Base-invariance implies Benford's law. *Proc. Amer. Math. Soc.*, 123:887–895, 1995. `http://www.ams.org/journals/proc/1995-123-03/S0002-9939-1995-1233974-8/`.

[6] Arno Berger and Theodore P. Hill. A basic theory of benford's law. *Probab. Surveys*, 8:1–126, 2011. `https://doi.org/10.1214/11-PS175`.

[7] Mark J. Nigrini. I've got your number. *Journal of Accountancy*, 5, 5 1999. `http://www.journalofaccountancy.com/issues/1999/may/nigrini.html`.