Ve401 Probabilistic Methods in Engineering

Spring 2018 Term Project 1

Date Due: 12:00 PM, Monday, the 2^{nd} of July 2018

JOINT INSTITUTE
交大密西根学院

General Information

The goal of this term project is to help you apply your new-found knowledge of probability theory and statistics in extended tasks that are beyond the scope of ordinary assignments. **It is strongly recommended that you do not leave the entire project to the last minute** but rather commence work on individual parts as soon as you are able to do so.

Group Work

You will be divided into groups of *4–5 students* each.

> Each group member must be familar 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 invovled 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. 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*.

Grading Policy

This term project accounts for 20% of the course grade; it will be scored based on

- **Form (4 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 (4 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 (12 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 receive the same grade for the term project. (Exceptions are possible in special circumstances.)

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 [3] 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 [1]. 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.

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 material constants, such as the Elastic Properties of Elements [6] and list the values of the shear modulus for the solid elements. Create a histogram of their frequencies and comment on the data.

iii) Recalculate the material constants in different units (e.g., pounds per square inch) and create an analogous histogram. Comment on your results.

iv) One convincing 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 [5] gave an argument that purported to show that this scale invariance implies Benford's law.

Work through the publication and restate his proof in your own words, filling in all the gaps (e.g, when he writes "Step 2 follows immediately from Step 1, and Step 3 is easily verified.", then explain *how* Step 2 follows from Step 1 and actually go ahead and verify Step 3. Furthermore, make sure that all terms you use are well-explained and that the end result becomes clear. You should also follow up on the relevant cited results in the article and at least sketch the ideas of their proofs. Note that the terminology in your sources may be a little old-fashioned; do not just copy it, but re-phrase it in the language you are used to.

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.

v) What does Benford's law predict for the second digit (which may be 0, of course)? How does the frequency of the digits behave in higher order decimal places? How are these statements derived?

vi) What are the shortcomings of Pinkham's approach? Why is it not a "full" proof of Benford's law? Theodore Hill [2] gave the first mathematically rigorous proof in 1995 and discusses the problems that previous attempts at proofs had. While you should not duplicate Hill's proof in the way you did Pinkham's argument above, summarize his basic ideas and discuss why Pinkham's proof does not give a complete explanation of Benford's law.

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 [4].

References

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

[2] 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/`.

[3] 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`.

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

[5] 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`.

[6] 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 29$^{\text{th}}$, 2015.