MONT105Q – Mathematical Journeys
Information on Final Projects
March 7, 2016

General Information

As announced in the course syllabus, one of the assignments for the seminar this semester will be a final project. You will be working on this project in *teams of 2 or 3* and the goals will be to prepare a roughly 10 page research paper and an oral presentation to the class on your project. The presentations will be given the final class meetings of the semester – May 2, 4, 6, and 9.

Schedule and Deadlines

- By Friday, March 18 – inform me which general topic you want to work on and who you will be working with. If your proposed topic has several different directions that might be pursued (see the descriptions below), please give an indication of which aspect you would like to work on. If you need assistance in forming "teams," I will be happy to help with that. Ideally, each group will work on a different project, although in some cases, the topics are large enough that if more than one group wants to try that, there will be ways to "split up" the topic into several parts. See me to discuss the possibilities.

- Friday, April 15 – Each group will submit, *by email*, a bibliography of sources to be used for your project. You should identify *at least six articles, books, or web sites*, that will be relevant. For each of your sources, write up a short paragraph giving a rough description of how that source relates to your main topic, what kind of information you will take from it, and how you will be using it (including a preliminary estimate of how reliable you think the information there is). The project descriptions below contain some first places to look, but you should plan to spend some time searching for additional sources of information. *Your final project papers may also refer to additional sources if you find that is necessary.*

- During the week of April 18, each team will meet with me during office hours (or at another time if that is not convenient) for a progress report and a chance to look at any questions that have come up as you have started to work on the project. I will be happy to discuss any aspect of the project at other times too, of course.

- We will decide which groups present which day when we get closer to the dates.

- All final project papers will be due no later than 5:00pm on *Monday, May 9*. (This is an absolutely firm deadline.)

Other Information

- Ms. Barbara Merolli, our Science Librarian, will be visiting our class after Easter Break to introduce herself and give some introductory information about using the library

resources to identify sources for your project. She will be more than happy to assist you in the important process of assembling the resources you will use for your project.

- The presentations can be done with PowerPoint. I will be happy to help out with technical details if needed.

- I will ask each group to do a "dry run" of their presentation with me at least one day before you go in front of the class. The purpose of this is to give you some feedback about what is working and what is not, and to give you some practice to minimize the effect of "nerves" when the time comes for the real thing.

The Final Project Report

This writing assignment will be somewhat different from the papers that we have done previously in this class. You should think of the written portion as the final write-up of the investigations you did on the data sets you looked at. So you should probably *not* try to start writing until most or all of the mathematical/statistical work has been done and you have thought over what the results said carefully. The evaluation of your project reports will be based on how well you have addressed the following guidelines and expectations:

- Distill your investigations into a central argument. A good "technical report" of this kind should be more than just a compilation of all the different things you did. It should be *well organized* and the writing should give the reader a clear indication where you are heading with your central argument at all times.

- A typical outline/breakdown into sections might look like this:

  - *Introduction* – Catch the reader's attention, explain the significance of the problem or topic. Say what you will do in general terms, without going into all the details from the start.
  - *Methodology* – Say where the data came from (and what the quantities are and what units they are measured in), describe the statistical methods you used and how you applied them.
  - *Results* – Say how what you found addresses the central argument. In discussing the results of statistical tests, it is considered good form to report the p-value returned by the test (the smallest α for which the null hypothesis would be rejected) as an indication of the strength of the conclusion. However, it is not necessary to reproduce the calculation of the test statistic, and the mechanics of carrying out the test, etc. Also, don't go overboard by including every single table and graph you generate. Be selective – one well-chosen graph can be just as informative as several similar ones! On the other hand, *don't "cherry-pick" your results* – be honest and include findings that might point to a limitation in the methods you used, or a shortcoming of your main argument.

  - *Discussion* – Don't overstate the importance of your findings, and (again) be honest if there are limitations. Give proper credit to sources you consulted that contributed to your ideas about the problem you studied. (In a more formal article, it would be expected that another section reviewing the most relevant contributions of previous work on related subjects would be included – that kind of full literature review is *not expected for this assignment*.) Discuss, if possible, how your results could be extended or generalized.
  - *References* – Include all books, articles, websites you used in the preparation of the work. For the websites, give the full URL, and the date you consulted.

- Be clear, concise, and correct in your writing. Aim for *no typos, misspellings, or grammatical problems*. But even more importantly, each paragraph should have a clearly evident purpose in relation to your main argument.

- Use graphs, tables, of data sparingly in the main text. (If you want to include more of these, that can be done in an additional Appendix section at the end.)

- Proofread your work carefully and have an "impartial" reader or readers look at it and give you comments. This can be one of the other teams or me. Be prepared and willing to *revise* your work based on the comments you get. Of course, this means that the writing must not be put off until the evening of May 8(!) Be sure you get started early enough so that the input can be put to productive use.

Ideas for Project Topics

The first group of project ideas use statistics to understand trends and patterns in data *that change in some way over time*. The technical name for this sort of data set is a *time series*. For example, a time series might represent average yearly temperature or precipitation at a particular location, or the batting average of a baseball player over the years of his career, or the concentration of a chemical in a water source measured in successive weeks.

Statistical Background

We can think of breaking a time series up into several components:

$$\text{Time Series} = \text{Trend} + \text{Cycle} + \text{Residual},$$

where the Trend might be an upward or downward movement, the Cycle might represent a regular, repeating changing pattern (e.g. the normal temperature changes due to the change of the seasons), and the Residual represents random variation. One of the most important questions to ask in dealing with time series data is whether there is a way to identify whether there is some increasing or decreasing long-term trend involved in a given time series.

There are many different methods that statisticians use to address questions of this type. *For examples and further discussion, it will be valuable to start out by identifying and studying a general discussion of this topic that is not too technical (i.e. geared toward people who actually need to perform such analyses rather than toward statisticians who are interested in extending these methods or developing more powerful ones.*

One basic method would be to use a *regression* of the data versus time. The slope of the regression line can then be used to decide if there is an upward or downward trend over time. However, if the trend is not linear, or if some of the necessary conditions ("football-shaped scatter plot" or homoscedasticity, normal distribution of the residual term, etc.) for the regression analysis are not met, then this approach is not especially appropriate.

There is an alternative class of methods called *nonparametric statistical methods*, including in particular a method called the *Mann-Kendall test for trends*. These do not require any of the assumptions needed for regression. As a result, the Mann-Kendall test has been very widely applied to study time series arising in areas such as pollution control, climate science, and other areas. The Mann-Kendall test works like this. Call the time series x_i, for $i = 1, \ldots, n$.

- Essentially the only assumptions necessary (for the basic version) are (1) if there is a trend, then it is "monotone" (either increasing for the whole time period, or decreasing for the whole time period), (2) there is no nonzero periodic Cycle term, and (3) the Residual term is purely random (not "autocorrelated").

- The test proceeds by computing for each time i and all later times $j > i$ the *sign* of the difference $x_j - x_i$ ($+1$ if $x_j > x_i$, -1 if $x_j < x_i$, and 0 if $x_j = x_i$). Let S be the *sum* of all of these signs.

- Then a quantity called the *variance*, V, is computed by the following formula:

$$V = \frac{1}{18}\left(n(n-1)(2n+5) - \sum_{k=1}^{g} t_k(t_k-1)(2t_k+5)\right),$$

  where g is the number of different groups of "ties" in the time series data, and t_k is the number of terms x_i in the kth group of "ties" for $1 \le k \le g$. (If there are no duplicate values in the time series, then this last term is zero.)

- Then the statistic

$$Z = \begin{cases} \frac{S-1}{\sqrt{V}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sqrt{V}} & \text{if } S < 0 \end{cases}$$

  is computed and this is used to infer the presence or absence of trends as follows.

- Provided that $n > 10$, Z is approximately *normally distributed*, so we can set up confidence intervals or hypothesis tests with rejection regions specified by the percentage points of the standard normal curve.

- For instance, for a (upper-tail) test of the alternative hypothesis H_a: There is an increasing trend in the data, versus the null hypothesis H_0: there is no upward trend, at the $\alpha = 5\%$ (Type I error probability) level, we would reject H_0 if $Z \ge 1.645$, and

not reject H_0 if $Z < 1.645$. As usual with hypothesis testing, the rationale for why the test is set up this way is that the chance that Z has a value greater than 1.645 is only 5% if the null hypothesis is true. There are corresponding two-tail tests as well, where the alternative hypothesis would be that there is some trend (either upward or downward).

- For smaller length time series, $n \leq 10$, there are tables available in many books and online that replace the standard normal table for setting up the corresponding rejection regions.

Among the advantages of the Mann-Kendall test are that it applies very widely, its conclusions are not greatly affected by gross errors or outliers (because it is only whether an increase or a decrease has occurred from one time period to another that matters, not the *magnitude* of the change), and the computations can be carried out by hand if necessary. One disadvantage is that it does not apply (and generates misleading results) when an upward or downward trend is combined with a cyclic (seasonal) variation. There are "souped-up" versions that deal with seasonal variations as well, though ("Seasonal Mann-Kendall" tests). Other "corrections" have been devised to deal with cases where the Residual term is not purely random.

A Mann-Kendall Example ("By Hand")

Say we have a time series with successive values:

$$12.9, 14.8, 13.2, 15.7, 13.8, 15.6, 12.9, 11.0, 17.3, 15.9, 19.0, 18.3, 20.1, 21.7$$

The Mann-Kendall S is computed by comparing each value *to all later values* (not any earlier ones) and recording a 1 if the later value is larger, a -1 if the later value is smaller, and a 0 if the later value is the same: Here we get the following

	14.8	13.2	15.7	13.8	15.6	12.9	11.0	17.3	15.9	19.0	18.3	20.1	21.7
12.9	1	1	1	1	1	0	−1	1	1	1	1	1	1
14.8		−1	1	−1	1	−1	−1	1	1	1	1	1	1
13.2			1	1	1	−1	−1	1	1	1	1	1	1
15.7				−1	−1	−1	−1	1	1	1	1	1	1
13.8					1	−1	−1	1	1	1	1	1	1
15.6						−1	−1	1	1	1	1	1	1
12.9							−1	1	1	1	1	1	1
11.0								1	1	1	1	1	1
17.3									−1	1	1	1	1
15.9										1	1	1	1
19.0											−1	1	1
18.3												1	1
20.1													1

Counting up, there are 91 pairs, 18 -1's, 1 0, and 72 1's. Hence

$$S = -18 + 0 + 72 = 54$$

The variance is computed like this (there is one pair of equal values):

$$V = (14)(13)(33)/18 - (2)(1)(9)/18$$

Then the Mann-Kendall statistic is

$$Z = \frac{S-1}{\sqrt{V}} \doteq 2.9058 \quad \text{(approximately)}$$

Interpretation: There appears to be an upward trend in the data and the Mann-Kendall test confirms this with a pretty small p-value. The probability of observing this large a value of Z if there were no trend would be about $p = .0018$ (using the standard normal table).

As indicated above, it is certainly *possible* to compute the Mann-Kendall Z-statistic by hand if necessary. However, as you can probably guess, for long time series this can be somewhat tedious and prone to computational errors if you are not careful. For that reason, it is much more common to perform the calculations in software. Many of the major commercial and research statistical software packages contain commands or have add-on packages to do this computation. There is also an Excel template spreadsheet called MAKESENS (developed in Finland for environmental applications) that is set up to perform these calculations. This can be freely downloaded from various sources on the web, including:

`http://en.ilmatieteenlaitos.fi/makesens`

There is a users' manual for use of this spreadsheet available here too.

Important Notes:

- If you use this spreadsheet, you will need to modify the Annual Data input page to accept different input time series – the Calculate Trend Statistics "button" should then compute everything you need.
- Because this spreadsheet involves macros – small programs embedded in the cells of the spreadsheet – you will probably need to override security features of Windows, your antivirus software, etc. when you download it to get a working version. See me early if you want to use this so that we can get the technical issues resolved.

Ideas for project topics related to trend analysis

Topic 1. Implementing the Mann-Kendall and Seasonal Mann-Kendall Tests.

Even though there are available Mann-Kendall spreadsheets and packages available for general use like the MAKESENS Excel template mentioned above, I am still a firm believer that when you are learning a new computational process, then it can be very valuable to "convince" a computer to do it for you (;)) by programming the process (either in a spreadsheet, or in some other sort of programming environment). For this project, the goal would be to develop your own Mann-Kendall and Seasonal Mann-Kendall procedures and test them thoroughly on various inputs. Everyone else would be using the Mann-Kendall

test more or less as a statistical "black box" – your presentation would involve digging a bit deeper into exactly how the tests work and explaining some of the fine points. Note: If you choose to work on this one, you should have a fairly high level of skill and experience in either spreadsheet macro programming, or in programming in some other environment (e.g. C++, Maple, etc.) If you have *never* done anything like this, it would be better to consider a different topic.

Topic 2. Analyzing Trends in Climate Data

There are a large collection of interesting data sets related to measurements of levels of various "greenhouse" gasses such as CO_2, NO_2 etc. in the atmosphere, temperatures, rainfall amounts, snow cover levels and durations, and many other things available for download from the web site of the Carbon Dioxide Information Analysis Center (CDIAC) at Oak Ridge National Labs,

`http://cdiac.ornl.gov/`

and the National Atmospheric and Oceanographic Administration (NOAA) National Centers for Environmental Information (NCEI):

`http://www.ncdc.noaa.gov/climate-information/statistical-weather-and-climate-information`

Some of these data sets are really *too simple* for the kind of trend analysis provided by regression or Mann-Kendall. For instance if you look at the time series of atmospheric CO_2 levels measured at various locations on the Earth, there is a clear monotonically increasing trend (superimposed on a seasonal variation that is tied to the growth cycles of plants) that requires no statistical analysis whatsoever(!) On the other hand, whether there are trends for some of the other trace gas amounts (e.g. carbon monoxide, chloroform, carbon tetrachloride, chlorofluorocarbons, etc.) is far less obvious. One very good project here would be to "pick your favorite trace gas," look at the various data sets available here and perform trend analyses, comparing results from different locations. (*Note:* There is certainly enough to do for 2 or 3 teams to work in this general area.)

One interesting question is how precipitation amounts have changed at the locations where temperatures have increased over the past 50 or so years. Do you find significant associations? (*Note:* Again by looking at different geographical areas, 2 or three teams could work on different aspects of this.)

Topic 3. Analysis of Trends in Baseball Statistics

This is literally a *huge* topic, as you can probably guess. Two or three groups could easily find plenty to do on various aspects of this. I recommend using the

`http://www.baseball-reference.com`

web site as your primary source – this has team and individual stats for almost the whole history of professional baseball(!) Among the kinds of questions that might be interesting to look at are: Are there typical patterns for statistics like yearly home run production, or on-base percentage, or slugging percentage, or OPS (on-base plus slugging) for individual

batters, or statistics such as earned run average, strikeout rate per nine innings, etc. for pitchers over their careers? Does this make it easier to understand some personnel decisions that team managements make? On the other hand, you could ask how the game itself has changed over time by looking at aggregate (average) home runs, slugging percentage, earned run average, etc. over some range of years. Then again, you could look at how the use of particular tactics like the base stealing, sacrifice bunts, etc. has changed over time.

Another sort of question: Are professional baseball players getting better? Are they getting more consistent? Is there more or less variation in the level of achievement of all batters now versus 50 years ago? Similarly, you could ask about the variation in pitching statistics! I'm purposely going to leave this one very open-ended because I'm very interested in seeing what you will come up with!

The next topics don't necessarily involve trend analysis, but they could be approached with some of the same ideas about time series data described above.

Topic 4. The "Hot Hand Phenomenon" – Fallacy or Reality?

In 1985, Thomas Gilovich, Robert Vallone and Amos Tversky published an article [GVT] called "The Hot Hand in Basketball: On the Misperception of Random Sequences." Contrary to the testimony of many basketball players and sports fans that there were times when players were "in the flow" or "hot" and were both more likely than not to make their next shots and aware that they were playing at a higher level, the authors of this article claimed to show via a statistical analysis that the data just didn't support these perceptions. In fact, they said, the outcomes of previous shots seemed to have no effect at all on whether subsequent shots were made or not. Charles Wheelan mentions this study on pages 102-103 of *Naked Statistics* as an example of when we misinterpret independent events as being *not independent*. This article attracted a lot of attention and has generated a fairly substantial literature about the "hot hand" phenomenon from the psychological point of view.

However, in September 2015, Joshua Miller and Adam Sanjurjo released a preprint "Surprised by the Gambler's and Hot Hand Fallacies? A Truth in the Law of Small Numbers" in which they claimed to have identified a flaw in the [GVT] analysis. When that flaw is corrected, they claimed that the data show there really can be "hot" (or "cold") hands in this sort of situation. Slightly later, Brett Green and Jeffrey Zwiebel, of the Stanford University School of Business released another preprint called "The Hot Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments? Evidence from Baseball" that claimed to find results similar to Miller and Sanjurjo's in baseball data. A good place to start looking at all of this is an article on `slate.com` by Jordan Ellenberg describing this rethinking of the [GVT] results.

I think it's probably safe to say that the jury is really still out on all of this, but it's an interesting question for any sports fan, and a really strong indication of the fact that the application of mathematics and statistics to real world questions is more like the rest of science (where new insights can lead to radical rethinking of our basic understanding of how the world works) than a realm of pure mathematical, Euclidean certainty.

There are several possible projects here:

- A (much) more theoretical topic would be to look closely at the original arguments and

data in [GVT] and try to understand why Gilovich, Vallone and Tversky came to their original conclusion. Then try to understand the criticism that Miller and Sanjurjo raised and why they think [GVT] "got it wrong." There's some pretty sophisticated mathematics here, but it will be interesting to work through this if you're up for a challenge.

- It would also be possible to look at new or different data (different sports, for instance) to see whether the hypothesis of the "hot hand" seems to hold up there. You should try to follow the sorts of analysis that [GVT] and Miller and Sanjurjo did here.

- It would also be possible to design a questionnaire and carry out a survey of students at the college to see if they have heard of this new research, and determine whether they believe in and/or think they have observed or experienced the "hot hand." Then you would need to analyze the data from that survey. For instance, it would be interesting to know whether there is a difference in perceptions between varsity athletes and the general student body, whether there are differences in perception between different sports, etc. **Important Note:** If you want to try this, there is a procedure you will need to follow to get approval from the Human Subjects Committee for the research, and this will take time, so talk to me soon to get the process started.

Topic 5. A Topic of Your Choice

If there is another topic you would prefer to work on, I am open to suggestions. This could be a similar application of trend analysis to a different subject. Or, you might want to try working on something completely different (something possibly related to one of the readings this semester). If you want to propose a topic of your own, you *must get my approval* before starting to work. For the March 26 deadline above, write up a short description of the topic or questions you want to look at and how you want to try to address them. I will let you know as soon as possible whether you have my approval.