MONT 107N – Understanding Randomness Seminar
Final Projects
March 8, 2010

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* 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 full week of the semester – April 26, 28, and 30.

Schedule and Deadlines

- No later than March 26 – 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.
- During the week of April 12, 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.
- The presentations can be done either with overhead projector slides, or with PowerPoint. I will be happy to help out with technical details either way.
- 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.
- All final project papers will be due on Friday, April 30.

The Final Project Report

This writing assignment will be somewhat different from the "opinion 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 April 29(!) Be sure you get started early enough so that the input can be put to productive use.

Statistical Background

All of the project topics below deal with using statistics to understand trends and patterns in data *that change in some way over time.* The technical name for this kind of data is *time series data.* For example, a time series might represent the GNP of a national economy over a sequence of years, or the batting average of a baseball player over the years of his career, or the temperature at a particular location at a particular time of day over a range of days, or the concentration of a chemical in a water source measured in successive weeks.

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. Chapter 12 from the following publication from the U.S. Geological Survey (*Techniques of Water Resources Investigations of the USGS, Book 4, Hydrologic Analysis, Chapter A3 – Statistical Methods in Water Resources, by D. R. Helsel and R M. Hirsch* USGS, 2002) gives a good (but perhaps somewhat technical) overview of several approaches:

`http://pubs.usgs.gov/twri/twri4a3/html/pdf_new.html`

(geared, of course, toward applications in water resources management).

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 appropriate.

There is an alternative class of methods called *nonparametric statistical methods*, including in particular a method called the *Mann-Kendall test for trends*, that 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.
- 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 $1 \leq k \leq g$. (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 $Z \geq 1.645$, 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 even 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.

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 a commonly-used 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://projects.met.no/~emep/assessment/MAKESENS_1_0.xls`

There is an manual for use of this spreadsheet at

`http://www.fmi.fi/kuvat/MAKESENS_MANUAL.pdf`

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. (Note: Because this spreadsheet involves macros – small programs embedded in the cells of the spreadsheet – you may need to override security features of Windows, 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.)

Project Topics

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, 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. BASIC, C++, Maple, etc.) If you have *never* done anything like this, it might be better to consider a different topic.

Topic 2. Analyzing Trends in Water Resources Data

Chapter 12 of the USGS publication mentioned above has several very good exercises at the end dealing with analysis of real-world data from various water resources management questions. The data sets are either presented in the questions, or are available from the web page listed above in an Excel spreadsheet or a text data file format. For this project, you would basically develop solutions for one or two of those exercises by working with the data and Excel. (You might copy the given data into the MAKESENS template, for instance, to do the calculations.) Then analyze your results. (*Note*: It would be OK if more two group wanted to work on these data sets. If so, we would just need to split things up so that you were looking at different questions. Also, when the questions say "use all the methods presented in this chapter," it would be OK just to do a regression and then a Mann-Kendall analysis, and compare and contrast the conclusions.)

Topic 3. Analyzing Trends in Time Series on the Environment and Climate Change

This topic would be similar in spirit, but somewhat more open-ended than the previous one. 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/`

In a way some of these data sets are 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 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. The measurements from the CSIRO Gaslab flask sampling network are especially good here because there are 9 different locations that measured various gas levels monthly over the period from 1992 to 2001. (*Note:* There is certainly enough to do for 2 or 3 teams to work in this general area.)

For climate data, perhaps the most interesting collections of data sets for US climate data are on the United States Historical Climatology Network. (There is a link to here from the CDIAC web site above.) 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 4. 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 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!

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.