STAT 100 Chance and Data Analysis October 2, 2002(Rev)

Today: Continuing follow-up of Sept 27 survey
Sample Midterm

Next, we look at the implications of the stock market simulation (most of these notes were posted on the weekend but we did not go through them in class yet.)

The Random Walk Simulation and the Stock Market

The main point requiring clarification in this area is what the simulation implies about the stock market. The basic random walk we used was the so-called "symmetric" random walk, in which steps of +1 or -1 are taken according to the toss of a fair coin. The "symmetric" just refers to the fact that the probability of a head is 0.5. (When we discussed "interventions" in time series, we used some simulation where this probability was different from 0.5.)

Now suppose we have been "walking" for 100 steps, and the result so far is 60 heads and 40 tails, so we would be at +20 from where we started. The graph of this 100 steps would look like a trend upward. But from the fact that we are using a fair coin, or a computer simulation of a fair coin, we know the next step is just as likely to be -1 as +1. We have no reason to think that the upward trend would continue.

What does this imply about the stock market? Suppose you are considering buying a stock. You see that the last 100 days prices show an upward trend. If you believe that the past trend will continue, you might buy the stock hoping that you could later sell it at a higher price. But the simulation shows that it is possible for a time series to look like it has established a trend, so that a forecast of the future would have some reliability, even when the time series acts like a random walk. The apparent trend upward (or downward) could be an illusion. The historical pattern of stock prices might have no useful information for forecasting the future trend. So the investor might have to consider other reasons to invest in a company than simply to look at the trend of past prices.

Now it is possible for a company to have an increasing stock price for a good reason. Its earning prospects might be improving for example. This might well cause a trend upward in price, and the investor would be making a rational move to buy the stock. The point is that the information leading to the rational buy decision should not be based solely on the appearance of a trend in stock price. That is what the simulation has demonstrated.

Percentile and Percent

Predictably, some students are confused by the term "percentile" thinking that it is a percent. Just look again at the example I gave you in class, using the 105 assignment marks. There are 100 different percentiles that can be computed from the 105 marks. I showed how to compute the 40^{th} percentile. One just sorts the marks from low to high,

and march up the sorted list 40 percent of the way through the list, and where you stop will be at an assignment mark that happened to be 24 (out of 28). So 24 is the 40^{th} percentile. Note that 28 is not a percent, but the particular percentile you computed, the 40^{th} percentile, uses the number 40 in its calculation, and this 40 is a percent. So it is the thing that specifies which percentile you are talking about that is the percent. The percentile will be in the units of the data itself. (I think the source of confusion here was that I converted the 24 out of 28 into a percent, 86 percent, but this was a bad idea in this example!)

In class, another example was given using heights of a class of students. The 90^{th} percentile was guessed to be about 75 inches. "75 inches" is not a percent, the 90 is a percent. The percentile is in the units of the measurement, and this is true of the 50^{th} percentile, the 90^{th} percentile, or any percentile. So the percentile is 75, and the percent associated with this particular percentile is 90 percent. 90 percent of the heights are less than 75 inches.

Economic Indicators

The confusion here seems to be "what do I need to learn about economic indicators?". We have covered seasonal adjustment, which is one thing. The other major point is the practical one that economists use observations of certain time series to tell them what the economy is doing, so they can recommend to government policy makers what changes are necessary. In doing this, they need to know which time series are leading indicators and which are lagging. Leading or lagging what? The economy. A lot of economists and market watchers pay a lot of attention to these indicators, and it is a major source of time series that are discussed in the media every day. Statistical methods play a large role in summarizing and forecasting these time series and that is why they are included in this introductory course. The actual details of summarizing and forecasting these time series are beyond the scope of the course, except that you should have gained from the article some practice at "reading" the time series graphs.

Randomization and Random Sampling

Your dictionary will have several meanings for the word "random", all somewhat related but not exactly the same. A common theme is that random processes are unpredictable. But we use the word more specifically.

Randomization refers to the assignment of treatments to experimental units. The "experimental units" are just the things or people that are going to be used to form the comparison groups. A "Treatment" is the characteristic that is imposed on a group of experimental units in order to see the effect on the response. The assignment of treatments to experimental units is called randomization if each experimental unit has the same chance of getting any given treatment. For example, in the food study, the three diets were assigned to a group of three rats "at random". Each of the three rats has the same chance to be selected to receive the "C" diet, and similarly for the SH and LH diets.

Note that the assignment is such that the three rats end up with one C, one SH and one LH, and this is sometimes referred to as a random "allocation" process.

Random Sampling is random selection (whereas randomization was random allocation). This is the process of selecting a subgroup for measurement. Sometimes we refer to the big group as a "population" and the selected group as the "sample". The random sampling process guarantees that every possible sample of the same size has the same chance of being selected. This is the defining characteristic of random sampling.

Law of Averages

I am using the name "law of averages" to label the phenomenon that was explained in terms of a long sequence of fair coin tosses. The fact that the number of heads does not become close to the number of tails but that the proportion of heads does get close to 1/2. This just requires a little thought. Isn't it true that to get an equal number of heads and tails is more likely in 2 coin tosses than in 10 coin tosses. If that does not seem intuitively obvious, do the experiment! About 50% of the time you will get 1 head an 1 tail in two tosses, but the 5H and 5T will only happen about 25% of the time. If you tosses the coin 100 times, the chance of 50 H and 50 T is only 8%. Nevertheless, the proportion of heads will converge to 0.5. Numerically, you have
Proportion of heads = (number of heads/number of tosses)
and even though the number of heads can be quite far from n/2, the fraction of heads still will be close to 0.5.

Some students were wondering why I plotted (number of heads - one half the number of tosses). This is the quantity that would stay close to zero if the number of heads was close to the number of tails. I wanted to show that there was no tendency for the number of heads to get close to the number of tails, even if the coin is fair.

The big picture here is that what happens on average does not necessarily happen! There are not too many families with 2.5 children.

That covers most of the items students said were confusing them. Some of these were also things that students wanted to hear more about. Looking at the list of responses to question 3 of the survey, there were some other items to expand on.

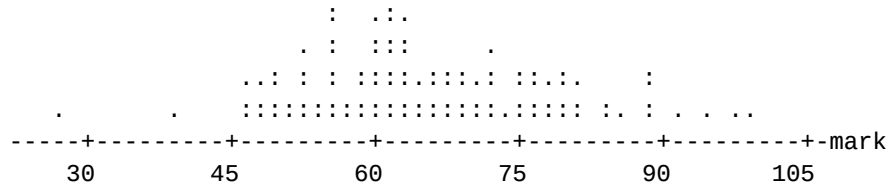
Mark adjustment, curving marks

Suppose you has list which related rank in class to the letter grade assigned. So if the marks for the course were based on 10 assignments out of 9 each, two midterms out of 45 each, and a final exam out of 120, then the maximum total score would be 90+45+45+120 =300. Suppose we decide that the top 20% of marks get A, the next 30% get B, the next 40% get C, and 5% get D and 5% get F (For simplicity, lets ignore + and - grades). If there were 110 students in the course, we would simply assign A grades to the top 20%, B grades to the next 30%, and so on.

Is this "mark adjustment"? Is this "curving the grades"?

It is a form of adjustment since if the exam were a little harder or easier than average, a rigid assignment of letter grades for certain mark ranges would be unfair.

It might be called curving the grades if it is based on drawing vertical threshold lines on the "curve" of the mark distribution. See the following graph:



If the letter grade is based on the mark distribution, and if the distribution is called a "curve" because of the shape of the above picture, then "curving the grades" is a possible way to describe the process.

More calculations, probability, formulas:

When the use of calculations and formulas helps to understand a statistical concept, I will provide it (as I have started to do in these notes). But there are some reasons to avoid this - some people do not like formulas and calculations, and they should not be required to use them if it is unnecessary to understand the concept. Secondly, students who rely on formulas can often fool themselves into thinking they understand the concept when they can do the arithmetic. Also, it is good practice to try to put a concept into words so that the concept has some real world utility - only a few people speak in symbols!

In the category of other things that students said they wanted (in question 4), a popular one was more detail about the midterm.

About the midterm and the nature of this course

The midterms and the final exam will be open book - you can bring notes or books to the exam. And a calculator. This should tell you something. Am I going to ask you to repeat things I have said in the notes? No. The questions will be ones you have not seen before, and so you will have to understand the concepts to answer the questions. The questions will be like the assignment questions with the constraint that they cannot be very time consuming since the midterm is only 50 minutes long. There has to be at least three questions or question parts so as to give some reasonable coverage of the material. I will give you an example of a midterm on Oct. 2, so you can see what I mean. But don’t make the mistake of expecting that the answers to the practice midterm will be useful in the real midterm. You should concentrate on the style of the questions to guide your studying, not the answers to the questions. And don’t expect to have time to look up the

answers in your notes - even if they were there you would not have time to look up everything. The open book is just to assure you that you don't have to memorize definitions, and to suggest that you need to understand and not just regurgitate what I have given you. Understanding is essential if the material is to be any use to you at all. No one is going to pay you a salary to calculate standard deviations or draw histograms unless you know when and how they are useful and what they imply. Chance and Data Analysis is a subject that includes far more than mere calculations, and this course is trying to introduce you to that subject.

I'll address some of the other points in class. The people who say the assignments are too hard or too long or too confusing, that the TA should give them the answers to the assignment questions, that the rhetorical questions in the notes should be answered, that the explanations should be simpler, ….

I would just say that a good course is not going to be easy, so prepared to struggle a bit to understand the ideas. There are a lot of them, because this is a good course!

Here is an example midterm test. It assumes you have read the articles on the Salk Polio Vaccine and on Health Insurance. It also assumes you have read the notes I post each day. Many of you will want to know the answers. In most cases, if you have the right answer, you will know it is right! If this confidence does not come to you, then you need to study the material more. It will not help you to learn the answer without studying the material! These questions will not appear on the midterm.

Remember to use the TAs in the Stat workshop (K9510) to help explain the course material.

STAT 100 Example Mid-Term I Oct 2, 2002

Instructions: This test is "open book" - you may use any books or notes. Attempt all questions. You have approximately 45 minutes for this test. Note there are 45 marks assigned to the questions. Point form answers are fine as long as they are comprehensible.

1. (15 marks)

The Salk Polio Vaccine Experiment used two different study designs in different geographic areas: the observed control approach and the placebo control approach.
a) What was the principal advantage of each design for providing the best information for the money spent?
b) Why was such a large scale experiment required, in which almost two million people were involved?
c) In the placebo control approach, what did the random allocation of people to treatment groups accomplish?

2. (15 marks)

In an assignment question you were asked to compute the SD of the average digit from the 7-digit serial number of a dollar bill. One way you could do this is by actually calculating the average serial digit of a few dollar bills, and then computing the SD of these averages. Describe one other strategy to find (or estimate) this SD of average serial number digit.

3. (15 marks)

A BC government study claims that 25 percent of new drivers have a traffic accident in the first twelve months of receiving the permit, based on data from the previous five years. In a study of current new drivers, in which a random sample of 500 new drivers' records are examined, only 100 had accidents. A commentator says that since the new study was based on only a sample of 500 new drivers, the new lower finding (of 100/500 or 20%) was not strong evidence that the rate had declined in BC. How would you use simulation to decide if the commentators conclusion was reasonable?

(KLW 021002)