STAT 100 Chance and Data Analysis Sept. 30, 2002

Today: Feedback and follow-up from Friday's survey

1. Most important concept since Sept 16.

| | FREQ'CY | |
|---|---|---|
| role of simulation in understanding real world | 18 | * * * * * * * * * * * * * * * * * * |
| role of SD in applications | 7 | * * * * * * * |
| risk and SD and diversification | 6 | * * * * * * |
| randomization | 4 | * * * * |
| seasonal adjustment | 3 | * * * |
| random sampling | 3 | * * * |
| role of means | 2 | * * |
| UV can mislead | 2 | * * |
| probability | 2 | * * |
| economic indicators | 2 | * * |
| chance and randomness | 2 | * * |
| calibration, curving marks | 2 | * * |
| percentile | 1 | * |
| data does not estimate probability perfectly | 1 | * |
| variability of averages | 1 | * |
| time series graphs | 1 | * |
| TOTAL | 57 | |

2. Most confusing concept since Sept 16.

| | | |
|---|---|---|
| seasonal adjustment | 11 | * * * * * * * * * * * |
| SD | 6 | * * * * * * |
| risk | 5 | * * * * * |
| variability of averages | 5 | * * * * * |
| stock market conclusion | 5 | * * * * * |
| percentile and percent | 4 | * * * * |
| economic indicators | 2 | * * |
| randomization | 1 | * |
| law of averages | 1 | * |
| smoothing method | 1 | * |
| practical use of SD | 1 | * |
| random walk with p≠.5 | 1 | * |
| variability and prediction | 1 | * |
| definitions | 1 | * |
| graphs | 1 | * |
| #H-n/2 | 1 | * |
| TOTAL | 47 | |

3. The topic you would like to hear more about.

| | | |
|---|---|---|
| stock market analysis,risk | 12 | * * * * * * * * * * * * |
| risk | 6 | * * * * * * |
| mark adjustment, curving | 3 | * * * |
| percentile | 3 | * * * |
| SD | 2 | * * |
| seasonal adjustment | 2 | * * |
| lottery calculations | 2 | * * |
| probabilities | 2 | * * |
| practical motivation for concepts | 2 | * * |
| after sept 25 | 2 | * * |
| calculations | 2 | * * |
| random sampling | 1 | * |
| gambling tips | 1 | * |
| daily life | 1 | * |
| motivation for experimental design | 1 | * |
| randomization | 1 | * |
| connections between concepts | 1 | * |
| simulations | 1 | * |
| South Africa | 1 | * |
| | 46 | |

4. Anything else you would like me to know.

| | | |
|---|---|---|
| details about MTI | 6 | * * * * * * |
| assignment questions confusing | 3 | * * * |
| want more office hours | 2 | * * |
| more definitions | 2 | * * |
| more examples | 2 | * * |
| want more math, fewer words | 2 | * * |
| more detail in notes | 1 | * |
| talk more slowly | 1 | * |
| emphasize important things | 1 | * |
| course grading scheme | 1 | * |
| assignments are a good review | 1 | * |
| MT should be easier than assignments | 1 | * |
| hard to relate simulation to real data | 1 | * |
| more intro to jargon of readings | 1 | * |
| too many unanswered questions in notes | 1 | * |
| want TA to answer problems like asst problems | 1 | * |
| simpler explanations | 1 | * |
| homework too long | 1 | * |
| | 29 | |

The 57 students who reported some important concepts that they had learned have a message for the student who said there was nothing he/she had learned: see the list of concepts it was possible to learn!

I would add that the list is ordered pretty well except that there is one idea that apparently did not appear as important as it should: **the variability of averages**. I will review this idea - it was a source of confusion for many.

But first, let me address the primary concerns in order of frequency:

**Seasonal Adjustment:**

This is not a major topic, but it should not be so difficult to grasp, so it is worth a few minutes of your time.

Many time series exhibit an annual pattern - like we saw with the gasoline consumption data and with the Vancouver rainfall data. The cause of this annual pattern is clearly something that has a similar annual pattern, such as temperature, precipitation, or the pattern of school activites. Often when one is examining such a time series, one is more interested in the variability that is NOT accounted for by the seasonal pattern. Seasonal adjustment is a way of adjusting the time series so that it reveals the departures from the usual seasonal pattern. The seasonally adjusted series might have certain variations that could be explained by causes that are not seasonal. For example, the gas consumption data, after seasonal adjustment, might reveal a steady decrease in miles per litre, possibly explained by an aging motor. Or, the Vancouver rainfall might have more extreme changes in precipitation than is usual for the seasonal effect, possibly revealing some el-nino effect.

How does one do the seasonal adjustment? There are refinements but the basic idea is to use smoothing, averaged over several years, to tell you the seasonal pattern, and then subtract this from the actual time series for the current period (of one or two years, say) to find the departures from the usual seasonal pattern. That is what we did. The result of this subtraction is the seasonally adjusted series. (Actually, one usually adds back in the mean response so that the seasonally adjusted series has numbers on the same scale as the original time series.) The seasonally adjusted series can then be examined for hints of causal influences that are *other than* seasonal.

**The Standard Deviation (SD):**

The SD is a measure of spread, or variability, in a set of numbers. Its value should be thought of as a "typical" value of the distance from any one of the numbers to the mean of the numbers - the size of a typical deviation. Why do we need this? Because it helps to summarize a set of numbers, along with the mean, and an important aspect of data analysis is summarizing data sets. If we know the average mark on assignment 1 was 83% and the SD was 16%, this tells in a compact way quite a bit about the distribution of

marks in assignment 1. This summary is especially useful if we want to compare two sets of numbers, such as assignment 1 vs assignment 2. More on this later.

For those of you who want a more mathematical description, it is the root mean square deviation from the mean. You don't need this formula to compute the SD, since the procedure I showed you is fairly easy to reproduce, but for those who want it, the formula is $\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \Big/ n}$ . The $\bar{x}$ is the mean of the n numbers $x_i$ . If you are using a calculator to compute the SD, you might get a slightly different answer since it is common to use n-1 instead of n as the divisor in the formula. However, the n-version is easier to understand and remember and the justification of the n-1 is very weak - so I choose to use the n-formula.

I have given a few demos of the calculation so that should suffice:

If you want another example to check that you can do it, try this:

SD of {0,1,2,3,4,5,6,7,8,9} is ? ANS is 2.87

If your calculator gives you 3.03, do not worry - it is using the n-1 formula. That is close enough! (To get 2.87, you have to multiply the 3.03 by $\sqrt{\frac{9}{10}}$ and in general by $\sqrt{\frac{n-1}{n}}$ .)

I have mentioned the general use, the definition, and the calculation procedure, of the SD.

One of the special uses of the SD relates to the variability of averages, and I will treat that as a separate topic:

**Variability of Averages:**

This is a very important application of the SD, hard to grasp at first, but very easy to use once you have mastered the idea.

An average is computed from a set of n numbers. We had an example in which you had a stock portfolio of five stocks which cost you $1.00 each, and the return on the five stocks in one year might have been, for example, `$0.50, $0.00, $0.50, $4.00, and $1.00,` then your average (or mean return) would be $1.20. Your total investment of $5.00 yielded a return of $6.00 - another way to say this is that your average cost was $1.00 and your average return was $1.20, or a 20% gain. 20% is a pretty good net return, but if you are considering doing the same thing for a second year, it would be nice to know how variable the return might be: Is a loss likely and if so how much? Will I have a chance to do better than 20%? These questions can be answered in part if you know the variability, for year to year, of the average return. But so far all you have is one year's experience with five of these companies. Is that enough information to estimate the variability of the average return next year?

There are two cases:

CASE 1: No prior info about risky-ness of companies

The data from one year is all you have, and you are willing to assume that the probabilities (assumed to be unknown this case) will be the same in year 2 as they are in year 1. In this case we use the SD of the observed data {$0.50, $0.00, $0.50, $4.00, and $1.00} to measure the variability that the unknown chance mechanism will display next year. This SD can be computed to be 1.44.
Here is where the theory comes in: The SD of the **average** yearly return (of the 5-stock portfolio) is estimated to be the estimated SD of observed yearly returns divided by the square root of 5, which is this case in $1.44/\sqrt{5}$ =.65. So we would estimate that the average return next year would have a variability of $0.65. In other words, our average return of $1.20 occurred in a situation where the other possible average returns might well be $1.20 +- $0.65 or even more extreme. Keeping in mind that $0.65 is a typical deviation and that deviations even twice as big might well occur, we could contemplate the average return being anywhere from a - $0.10 to + $2.50. (-.10=1.20-2*0.65, 2.50 = 1.20+2*0.65). In terms of average net return, that would be anywhere from a loss of $1.10 to a gain of $1.50.

This is useful information to the investor. Note however that the information is based on an estimate of the return SD determined from one years' experience, and this is not too precise an estimate - but it is better than nothing!

CASE 2: Prior Information about companies is available:

Now suppose the investor has the additional information that these companies have returns which occur according to a certain probability distribution. We specified this in our earlier discussion as

| Return | Probability |
|---|---|
| $0.00 | 0.25 |
| $0.50 | 0.25 |
| $1.00 | 0.25 |
| $4.00 | 0.25 |

Recall that we computed the average return to be $1.38. Now we also would like the SD of the average return, to again form a range like 1.38 +- something ….
Imagine a barrel with tickets on which are written 0,.5,1, or 4 and that there are the same large number of tickets for each number (0,0.5,1 or 4) in the barrel. We can think of the first years' experience as the result of drawing 5 tickets from the barrel. Similarly for the second years' experience. Although the particular tickets chosen will have various SDs, we can compute the SD of all the tickets in the barrel from a knowledge of their relative frequencies. The SD of all the tickets will be the same as the SD of the numbers 0,.5,1, and 4. (It is only the relative frequency that matter, and in this case the relative frequencies are all 25%). Now this SD can be computed to be 1.56. In this case we

would predict that the SD of the average return is $1.56/\sqrt{5}$ = .68. So we might say that the average return is $1.38 +- $0.68. Again, keeping in mind that the $0.68 is the typical deviation, we might contemplate average returns of from $0.02 up to $2.74, or in other words average net returns of -$0.98 to $1.74.

Again, this is useful information for the investor. Note that the information should be right as long as the prior information about the companies is correct. This case involves a calculation rather than an estimation.

In both cases, the SD of the average is computed from the SD of the individual values divided by the square root of n.

For the formula people,

$$\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$$

Note what this implies - that averages are less variable than the things that are averaged, and the factor of reduction of variability is $\sqrt{n}$, where n is the number of things averaged.

Got that? OK, next is

**Risk**

Risk = Probability of Loss (I write this as Risk = P(loss))

Consider again our risky companies. Let assume we know the probabilities specified for those companies.

Which would be the least risky strategy?

$100 invested in one risky company, or
$1 invested in each of 100 companies?

The risky company will return 0 or .50 for a $1.00 investment 50% of the time, so P(loss) = 0.5 for one company. Risk for the first option is .50.

We investigated the second option using simulation, we found that P(loss) = 0 approximately.

This answers our question. But can we now consider the case of $20 invested in each of five companies? Isn't it true that P(loss) will be the same as if we invest $1 in each of five companies? Now we just figured out in the CASE 2 above that an average net return could be anywhere from a loss of 98 cents to a gain of $1.74. This does not tell us P(loss) but if we imagine what the shape of the net return distribution looks like, from our simulation of 100 such companies, then we might guess that P(loss) is about 20%.

Or we could look at the simulation experiment described on Sept 25 and see that the result of the simulation of many portfolios of 5 companies was that a loss occurred about 25% of the time. So for a 5 company portfolio, risk=P(loss)=0.25.

Note as we increase the number of companies in the portfolio from 1 to 5 to 100, the risk decreases from 50% to 25% to almost 0%. The distribution of net returns narrows as the number of company returns averaged increases. Again, this is the phenomenon that **averages are more stable than the things averaged**. (More precisely described by the square root law).

(The formula people might want the following - for independent $X_I$ s having the same variance, we can compute the square root law….

Variance of $X_1+X_2+X_3+….+X_n$ = sum of variances of $X_I$ = n variance of $X_I$

Variance of $1/n\ (X_1+X_2+X_3+….+X_n) = (1/n^2)$ n Var $X_I$ = (Var $X_I$)/n

Var $\bar{X}$ = (Var X)/n so SD $\bar{X}$ = SD X/$\sqrt{n}$ )

Next, we look at the implications of the stock market simulation:

**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 $90^{th}$ percentile. One just sorts the marks from low to high, and march up the sorted list 90 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 $90^{th}$ percentile. Note that 24 is not a percent, but the particular percentile you computed, the $90^{th}$ percentile, uses the number 90 in its calculation, and this 90 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!)

**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**

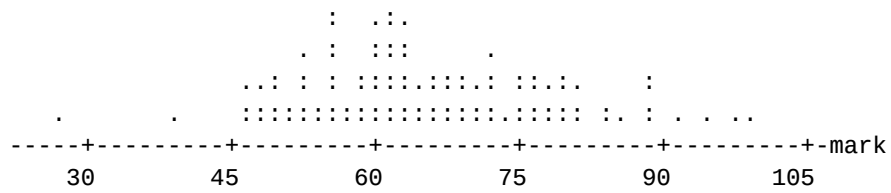
Suppose you has list which related rank in class to the letter grade assigned. So if the marks for the course were based on 10 assignments out of 9 each, two midterms out of 45

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 - eve 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!