STAT 100 Review Notes April 13, 2010

In addition to finishing a couple of topics from the April 8 lecture, we will cover as much as possible of the following in week 13.

5. (Probability Models - 59 choices) Normal, Poisson, and Gamma Distribution Models, Density, Variability, Models, and application to Sports Leagues.

5A. (Inserted April 7 – I missed this on the initial listing)
(Industrial Issues – 37 choices) Six Sigma, Quality Control, Variability Reduction, Reliability, Cell Phone Fraud.

6. (Smoothing - 31 choices) Moving Averages, Fuel Consumption, Survival Data, and Zipf's Law.

7. (Experiments and Observational Studies - 29 choices): Random assignment, randomization, Simpson's Paradox, and applications School Choice, Turkey Mail, Memory Load, Clinical Trials, and one I forgot to list - Gilbert Murder Case.

8. (Sampling Surveys - 17 choices): Political Opinion Polls, HIV study, Randomized Response, Veteran's Fund Raising, and Tiger Prey.

9. (Miscellaneous - 85 choices): Various details including Decision Errors, Sampling with and without replacement, SD or proportions, ...

Before I begin: Note a headline in Saturday's Vanc. Sun "The best forecast for the dollar is its current price". It is about the cost of a Canadian dollar in US currency, and it is described as having a pattern like a random walk – no surprise to STAT 100 students …

Probability Models

I have reviewed the idea that simulating samples from a probability distribution is logically the same as selecting a random sample (with replacement) from a population. (See Notes April 6, and April 8). We have discussed several phenomena by using the **N(0,1)** distribution as our population, and more recently the **Uniform** Distribution on (0,1) and the **Poisson** Distribution. In particular, we used the Normal "population" to produce data for the "theory of means" discussion even though it was not important to start with a Normal distribution. (To get away from the Normal population we used the uniform distribution on the integers {0,1,2,…9} to show that the theory still worked.) We used the Uniform Distribution to simulate a uniform spatial distribution of points on the unit square. We used the Poisson distribution to predict the number of empty cells from a uniform spatial distribution.

Another distribution that was "ad hoc" was the one that is uniform on {-1,-0.5,0,3}, and this was used to demonstrate the effect of diversification of risky investments. Yet another unnamed distribution was the one that was uniform on {-1,+1}, and this was used to describe coin tosses and a symmetric random walk. We also used a similar distribution – uniform on {Win, Lose} – to describe the outcome of a game between two teams of equal quality, in our simulation of sports league phenomena.

Yet another model that was mentioned was the **Gamma** distribution. Not much detail about this was discussed – only that it was a model for a population that had a density that had skewed shape. For example income distributions usually have a long right tail and so the Normal would not be appropriate, but the Gamma would be better in this application. The Gamma was also used without much comment as a population from which samples were selected to generate a distribution of sample means, to show that the distribution of sample means was still approx normal.

Density

One important idea that needs review is the idea of a probability "density". That bell curve I refer to often is a "density". So is the rectangular picture of the uniform distribution. To understand a "density" you need to realize that a density is a "function" – we usually write f(x) in mathematics for a function. If f(x) is a density then it provides a value of f(x) for each value of x, and f(x) is proportional to the relative frequency of the value x. The normal density is highest at its mean (0 in a standard normal) and falls off as x moves away from the mean. I have discussed how the SD is related to the shape of the normal density: the distance from the normal population mean to the point where f(x) changes its curvature, is 1SD. (1 unit for the standard normal).

The density for a population that contains all values in the interval (0,1) equally often is called the uniform density. Its f(x) is 0 when x is outside of (0,1) and equal to a constant value 1 inside that interval. If you understand "density" this should make sense.

Variability

A theme throughout the course is the presence of "variability" in data. This is one of those ideas that causes semantic problems. If our data values in a random sample from {-1,-.5,0,3} are (0,0,3,-1,3,-1,-.5,0,-.1,0), what do we mean when we say the sampling process produces variability in the sample? The numbers in the sample don't move around once recorded! What the word variability refers to in this context is the fact that the numbers in the sample are not all the same. We measure "variability" in the sample, or in the population, by computing the SD in each case. In fact we use the SD of the sample as an estimate of the SD in the population. In the sample shown it is 1.51, and in the population it is 1.80.

In other words, variability describes the outcome of a sampling process, not the change in the numbers once sampled.

This distinction is important in thinking about the sampling distribution of the sample mean. Although we usually have a single mean from our sample with which to estimate the population mean, we still talk about how variable the sample mean might be. It is the process of sampling that we are describing.

Models

Why do we need "models" like "Normal" and "Uniform" to describe populations, instead of just providing a list of population values. There are two reasons I will mention here:

i) Having easy ways to generate populations with known properties makes simulation very convenient, since simulation outcomes can be related to well-known population characteristics.
ii) Certain models are expected to occur in real data since the method of generating them leads to the model automatically (like normality of sample means, and Poisson distribution for uniform spatial scatter).

5A. (Industrial Issues – 37 choices) Six Sigma, Quality Control, Variability Reduction, Reliability, Cell Phone Fraud.

The words "Six Sigma" does not have any direct meaning in relation to the body of techniques that are implied by the modern use of the phrase. However, it is true that "Sigma" in statistics jargon is often used for the SD. As slight connection could be made by saying that the ±3SDs that include essentially all the "usual" data (in a normal distribution), but this is a stretch …

So what is "Six Sigma"? It is just a reasonable sequence of steps in solving industrial problems, especially those that involve data collection (and most problems do.) However, there are some useful techniques that help in this process: fishbone chart, run chart, control chart, pareto chart. You should know what these are and why they are useful.

I elaborated on the control chart in earlier lecture (march 30). The use of this was that it simplified the signal for a floor manager in a manufacturing plant (or similar institution) that something needed re-examination in order to eliminate sources of variation: management by exception.

Why is it important to eliminate sources of variation? A manufacturer can produce just what the product specifies and not more nor less – this is the condition for maximum profitability. See the March 30 lecture notes for more on this.

The washing machine reliability article (pp 339ff) and the cell phone fraud article (pp 293ff) both described industrial problems involving variability. The strategy in the washing machine article was to do an experiment involving accelerated testing to discover the source of reliability problems, and eliminate them so that sales could be increased. In the case of cell phones, variation was used to signal fraudulent use of a cell phone, and clearly elimination of this source of variation would also increase profits.

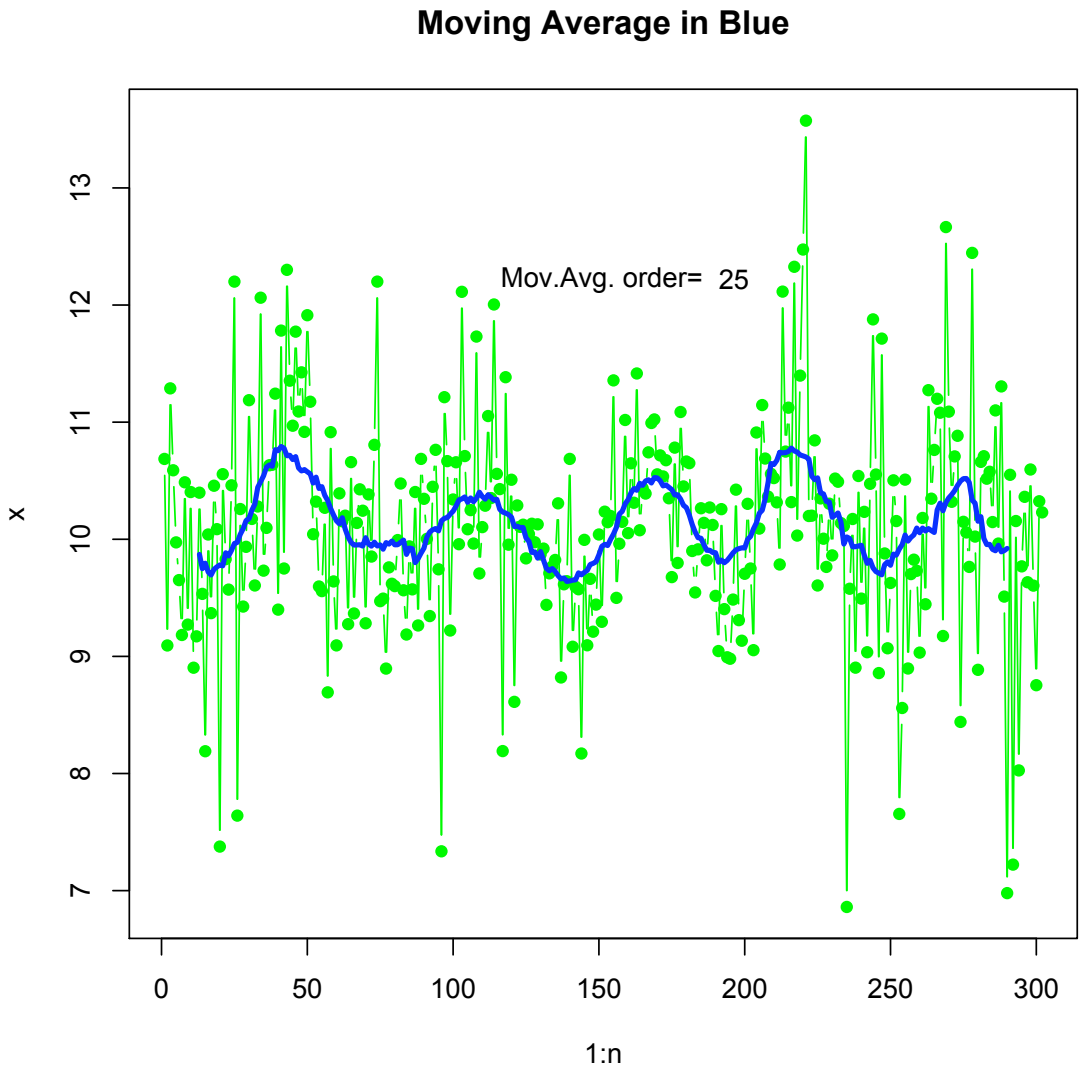
---

6. (Smoothing - 31 choices) Moving Averages, Fuel Consumption, Survival Data, and Zipf's Law.

Smoothing

This topic was partially reviewed already in the April 8 lecture, in connection with the description of dependence of time series. See those notes concerning "moving averages". Also, recall the following graph.

In the case of "Fuel Consumption" I did not use the moving average method for the smoothing, since it does not work well for high order moving averages – it chops off the beginning and end of the series. I used a method called "loess" which is outside of this course. Nevertheless, I can demonstrate how the moving average would work on that fuel consumption data.



So even the simple moving average (order 25 here) produces the regular sinusoidal pattern in phase with the annual season. With this large data set, the loss of the end points did not matter much (sometimes a problem with the moving average smoothing method.)

Other applications of smoothing: Survival Data (Car Accidents) and Zipf's Law (City Populations).

We used a straight line as a very rough smoothing of the survival data. As pointed out in the assignment, the straight line was not a very good model since it had to fail at the upper end (since the probability of having an accident for a certain exposure cannot be greater than 1). But the model was useful anyway.

The Zipf's Law graph showed that the rank of the city population multiplied by the

population itself was fairly constant, at least for Canada and the US. But it was not exactly constant. Again, the approximate model was useful for description and comparison, even though not quite right. We can think of the imposition of the straight line on the product values as a smoothing.

---

7. (Experiments and Observational Studies - 29 choices): Random assignment, randomization, Simpson's Paradox, and applications School Choice, Turkey Mail, Memory Load, Clinical Trials, and one I forgot to list - Gilbert Murder Case.

Experiments and Observational Studies

By now, I hope everyone knows the difference between an experiment and an observational study, and why that is a useful distinction: in an experiment the investigator assigns the treatment to be compared to the subjects (or experimental units) while in an observational study, the characteristic that defines the comparison groups is a characteristic of the subjects (or study units). The reason it is important is that only in an experiment is a direct inference of causality possible (the characteristic as a cause of the outcome), since in an observational study the characteristic of comparison might be associated with the real cause but not itself be the real cause.

For example, a study of people with high blood cholesterol to be compared with a group with low blood cholesterol might be assessed for subsequent heart problems, but even if the correlation seems solid, the likely inference that the cholesterol was the cause might be wrong since the high cholesterol itself might be associated with a high pressure job and the job stress might be the direct cause of the heart problems as well as the high cholesterol. Another group with high cholesterol but a low pressure job might have no such risk of heart problems. In this example, the job stress is a “lurking” that the investigator did not measure, but should have for a more useful analysis. And there could be many such variables – an observational study can never have comparison groups balanced with respect to all possible causes. On the other hand, an experiment in which the comparison groups have been randomly assigned are balanced (in a statistical sense) with respect to all “lurking” variables, and these alternative explanations of outcome differences of the comparison groups are not tenable.

The web site I referred you to about **Simpson’s Paradox** is a very interesting and clear explanation. Simpson’s paradox is just an illustration of how misleading an observational study can be, and a suggestion of how to minimize the likelihood of being misled.

The applications listed in the intro to this review section all involve expt vs obs

study issues.

School Choice (pp69 ff): Applicants were assigned at random to public/private schools
Turkey Mail (pp 373ff): Subjects and Day-of-Week were assigned at random to the emails.
Memory Load(pp 211ff): Subjects were assigned to the various orders of memory load sequences
Clinical Trials(Mar 4 notes and pp227 ff): Subjects are assigned at random to the treatment groups
Gilbert Murder Case (pp 3ff): an observational study in which randomness was rejected as an explanation of the prevalence of deaths on Gilbert's shifts, but Gilbert's guilt was still not proven by this result.

8. (Sampling Surveys - 17 choices): Political Opinion Polls, HIV study, Randomized Response, Veteran's Fund Raising, and Tiger Prey.

Sampling Surveys

Our discussion of Political Opinion Polls assumed a simplified scenario of basing preference for a Candidate on a simple random sample. The usual population sampled was assumed to be very large relative to the sample size, and the correction factor for estimating the SD of the finite population was so close to 1 that we could ignore it.

We showed by formula that once the population was more than about 20 times the sample size, the correction factor for sampling without replacement could be ignored, if for a finite population.

The difficulty of taking a simple random sample in the Canadian population was mentioned, since the list of all eligible voters was hard to access well before an election. Usually the sample in a political opinion poll involves random selection but is not a simple random sample. A similar difficulty arose in the HIV study – instead of sampling the target population directly, a list of venues was constructed and the venues were selected randomly – this worked since adjustments to the data gathered were made for the relative size of each venue.

The tiger prey "survey" was a quite different problem for researchers. Here they were simply trying to count deer in an area of known size, in order to estimated density and hence abundance. The problem was that the deer were very hard to find, not because they were rare, but just because they avoided being seen. The clue was the footprint data. However, the article describes how expert training was necessary to estimate density directly from footprints entering and leaving a certain area. The article shows that a simpler method requiring less expert

training was to count footprints crossing a certain linear path. Then the correlation between the linear density and the areal density could be used to estimate the real target which was the areal density. Regression was used for this. (Straight line regression only works when there is a straight line correlation, and the ordinary correlation coefficient is actually measuring the closeness of points to a straight line.)

The Veteran's Fund Raising study was a survey done serendipidously (data mining article p 307ff) as the previous year's fund raising event. Information was available on the last year's mailing list along with the response consequences of it. The sample examined was not a random sample – it included everyone on the mailing list. But the response certainly involved some randomness.

The last topic under this heading that students wanted a bit more about was the randomized response survey. This was just the simple use of probability to find out, in a non-confidential way, information from students that would usually be considered confidential. The method tries to ensure students that their response will not reveal their actual answer to the sensitive question, since most students answering the question in the affirmative will be answering a non-sensitive question (Did you toss a head?). Again, the randomness did not come from random selection, since the entire class present was questioned, but rather from the subsequent coin toss.

9. (Miscellaneous - 85 choices): Various details including Decision Errors, Sampling with and without replacement, SD or proportions, ...

I'll just reiterate a couple of things in this Miscellaneous Category.

Decision Errors

In many situations in which decisions are made on the basis of observed data, there is correct decision and an incorrect decision. Often these two decisions are made in each of two "states of nature": e.g. email was spam, or email was not spam. So there are four situations in each decision:

State of nature	Decide it is SPAM	Decide it is not SPAM
Spam	1	2
Not Spam	3	4

In situations 1 and 4, the decision is correct, but in situations 2, and 3, it is not correct. We sometimes refer to 3. As a False Positive, and 2. as a False Negative, especially when the decision is about the presence of disease. Statistics jargon also uses the unhelpful terminology "type I error" and "type II

error", but I have avoided this jargon in STAT 100.

SD of Proportions

The SD of sample proportions can be computed like any other SD, as long as the items to be analyzed are coded as 0s and 1s. So if the class consists of students {M,F,F,M,M,F,M,F,F,…,M,F} then to compute the proportion of Males, I would rewrite the data as {1,0,0,1,1,0,1,0,1,1,…,1,0} and then the proportion of Males would just be the proportion of 1s in the recoded data. So we might have the proportion of 1s as 56/140 = 0.40. But if we consider this class to be a random sample of SFU students (probably a bad assumption but we will do it anyway for illustrative purposes), then we realize that the sample proportion based on this class could have different with a different class, and we would like to know how close to the true proportion of males in our sample is to the same proportion in the SFU population. Suppose this SFU proportion is "*p*". We do not know *p* but we have our estimate of it $\hat{p}$=0.40.

What we need is the variability of $\hat{p}$. Can we estimate that? Noting that $\hat{p}$ is an average based on a sample of size 140, we actually do know how to estimate its variability: SD of population / √sample size.

What is the SD of the population? It can be shown that the ordinary formula for SD on the 0-1 data turns out to be $\sqrt{p(1-p)}$, where p is the proportion of 1s in the population sampled. But as I said before, we don't know p, but we do know the estimate of it, $\hat{p}$=0.40. So we can estimate the variability of $\hat{p}$ by

Est of SD of population / √sample size = $\sqrt{\hat{p}(1-\hat{p})}$ /√sample size = $\sqrt{.40(1-.40)}$ /√140 =0.041.

So a typical error in estimating p from our sample of 140 is about .041, since 0.041 is the estimated SD of $\hat{p}$.

We could say that an interval estimate of p is 0.40±.082, we would be correct, in the sense that the true SFU value would be in this interval, 95% of the time we use this method.

See the list of topics included in the final exam – posted on the STAT 100web page.

End of Course!

I wish everyone success on the final exam, and that in your careers you will look back at this course as having taught you something useful and interesting.

KLW 2010/04/13