STAT 100 Notes February 9, 2010

This week's lectures are about **Models.** There are several stories to read from the Peck reader, and each one has a context that requires a model and also some special considerations about how to make use of the model.

One model we already discussed is Zipf's Law. The importance of this is just that it is an easy example of a model that is useful but wrong. It is not a statistical law that needs to be memorized!

Other models arise in the following readings:

Gerow, Miquelle & Aramilev 105-118: Monitoring Tiger Prey Abundance in the Russian Far East.
Matis & Kiffe 119-134: Predicting the African Bee Invasion.
Madigan 135-148: Statistics and the War on Spam.
Brookmeyer 197-210: Modeling an Outbreak of Anthrax.
De Veaux & Edelstein 307-322: Reducing Junk Mail Using Data Mining Techniques.

The assignment due after the break for the Olympics involves these five readings, plus the reading from week 5 -
Eddy & Smykla 211-226: The Last Frontier: Understanding the Human Mind.

Assignment #6 (Due Tuesday, March 2, 2010)

1. (Eddy & Smykla 211-226) a) Why was the sequence of memory load conditions applied to each subject in a random order, as described in Table 1. (For example, why not just apply 0,1,2,3 three times in that order?)
   b) The strategies discussed on pp 217-219 were intended to reduce the variability of measurements taken under the same conditions. Why was reduction of variability important for achieving the objectives of the experimenter?

2. (Gerow, Miquelle & Aramilev 105-118) a) What does the interval (1.57, 1.95), produced on p 115, represent, and b) why is it useful for officials concerned with Amur tiger populations?

3. (Matis & Kiffe 119-134) a) To predict the time interval from the "last sighting" in Sept 1989 to the first US sighting, the authors use the data in Table 1 and eventually arrive at the distribution of time interval shown in Figure 4. In the inference from Table 1 to Figure 4, why was Figure 3 introduced?
   b) Figure 7 on page 130 provides information about the birth rate and death rate of the bee colonies population. Why does this graph imply that an equilibrium population level would eventually be reached?

4. (Madigan 135-148)
   What is the connection between the "bag or words" filter that Madigan eventually proposes, and the introductory example using only the word "free".

5. (Brookmeyer 197-210) Figure 3 shows the minimal data available at the time of the outbreak. Note that the times from the Florida exposure to the first cases were 5 and 6 days, and from the Washington exposure were 4 days. Why was it decided to use a mean incubation period of eleven days?

6. (De Veaux & Edelstein 307-322)
   a) How is logistic regression different from linear regression, and b) why was logistic regression used by the successful data miners for the PVA fund-raising problem?

Eddy & Smykla 211-226: The Last Frontier: Understanding the Human Mind.
Gerow, Miquelle & Aramilev 105-118: Monitoring Tiger Prey Abundance in the Russian Far East.
Matis & Kiffe 119-134: Predicting the African Bee Invasion.
Madigan 135-148: Statistics and the War on Spam.
Brookmeyer 197-210: Modeling an Outbreak of Anthrax.
De Veaux & Edelstein 307-322: Reducing Junk Mail Using Data Mining Techniques.

The assignment tries to draw your attention to some statistical issues in each story. But the assignment questions are not the only ones you need to think about – it will be necessary to read the articles.

Eddy & Smykla 211-226: The Last Frontier: Understanding the Human Mind.
(this note repeated from Feb 4, except addition of Simpson's Paradox.)

MRI – method of detecting map of brain activity
Investigator wanted to be able to detect a specific kind and intensity of brain activity
Devised a series of tasks requiring an increasing "memory load"
Wanted to show that increasing memory load could be seen as increasing activity in a certain part of the brain.

Design issues: What memory tasks to set
How to establish cause and effect – (expt, dose-response)
How to reduce variability of the response

"Use repetition, control what you can, and randomize the rest"
10 subjects
within-subject comparisons
repetitions within subjects

randomize the order of memory load intensities p 219

Result: Graph p 224

What do you need to get out of the article?

Principles of Study Design
Use of randomization
Importance of understanding context to analyze data

Turkey Mail (Kahn and Roseman article p 373) was an experiment. The units of study were the emails, and the "treatments" were the assignment of SUBJECT and DAY-of-WEEK.

Review of "Experiment": Clinical Trial is an example.=

Dangers of Observational Studies: Simpsons Paradox.

See the great examples at
See the wonderful examples at en.wikipedia.org/wiki/Simpson's_paradox

You might think that Simpson'e paradox is a weird curiosity that occurs rarely in real life. However, it is actually an extreme version of a very common effect – the effect of "lurking" variables in observational studies. When we compare some groups using one or two variables, there is often a third unrecorded variable that is unbalanced in the comparison groups and does have an influence on the observed variables. This can lead to erroneous inferences from the comparison. Simpson's paradox is just an example of this where the error is particularly startling.

Summary: True experiments require a lot of work and a context that is feasible from both a cost and an ethical point of view. When ethics are not a problem, good design can keep costs down while still providing the information sought (in spite of random variation).

Gerow, Miquelle & Aramilev 105-118: Monitoring Tiger Prey Abundance in the Russian Far East.

See page 107 – map. Khabarovsk, 600,000 population. Amur River. Latitude 49N
8500 km east of Moscow!
Deer – prey of Amur tiger – temperate forest – cold winters. Hard to count either tiger or prey. Count fresh tracks instead. But not easy.

Capture, Tag, Recapture Method – does not work
Tracks in and out of a small area – hard to do.
What if just count rate for all tracks – easy. Number of tracks per km.
But how to relate to density?
Density = number of animals per square kilometer.
See Figure 5. Note (0,0). D=C*T
In regression, like equal SD of D across horizontal variable T.

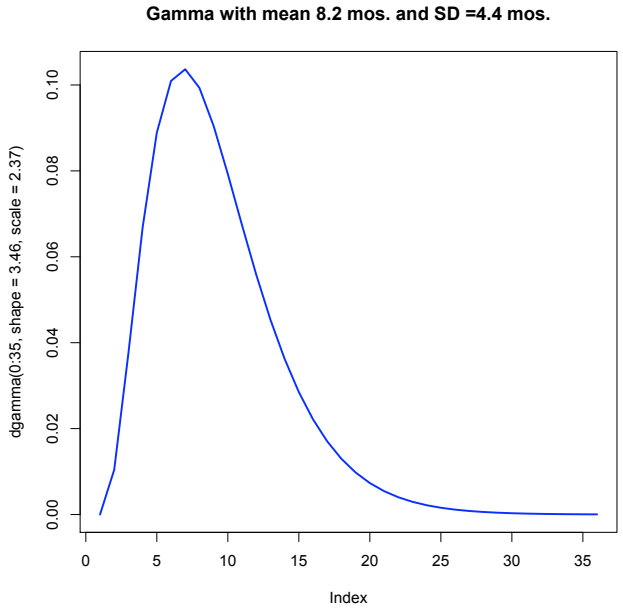
In this case log(D) does it. Note log(D) = log(C) + log(T)
If T and C known, can compute D.
For data where both T and D are known, can estimate log(C)
From article data, log(C) = .56 and SD of estimate of log(C) is .055.
Convert to C in interval 1.57 to 1.95. Can use to convert T estimate to D estimate.

Summary: The difficulty of counting forest animals led researchers to use an indirect method based on a statistical link: count tracks to predict density.

Matis & Kiffe 119-134: Predicting the African Bee Invasion.

Fig 1 shows progression
Table 1 has "speed" data between surveys.
km/mo. was speed. but used mo./100km = "transit time"
Model for transit time used gamma distribution. Common for times.

Gamma distribution: a model like the normal – but it is not symmetrical, it has a "right skewed" distribution meaning a long right tail.



Gamma has parameters alpha and beta but it is possible to express these in terms of the mean and SD. Mean = alpha * beta and SD =alpha^0.5 * beta. The reason alpha and beta are used to describe the gamma distribution is that alpha and beta control separate features – alpha is a shape parameter and beta is a scale parameter.

This distribution describes the transit time (in months) for the Africanized bee invasion to traverse 215 km. Knowing the mean 8.2 months gives a point estimate, but knowing the SD also gives a way of describing the possible error in the 8.2 prediction. Looking at the graph we can see that transit times of 3 months or 20 months would not be too unusual.

Another statistical strategy introduced in this article is the "logistic growth curve". This is a particular mathematical description of a typical population growth: slow start, then maximum growth, then slow down as the population reaches the capacity of the growth environment – often called an S curve. An explanation for why this happens is given by the birth and death rate curves on p 130. The difference between the birth rate and the death rate gives the rate of increase of the population.

Summary: The article describes a natural phenomena for which statistical tools provided a way to describe how the phenomena would unfold. Some simple graphs were used to describe the result of the analysis, but the analysis itself involved some mathematical models: gamma distribution and a birth-death process.

Madigan 135-148: Statistics and the War on Spam.

The idea at the beginning of the article is that an email with a word like "Free" in it is more likely to be spam than one that does not. But obviously this is an imperfect filter. There are two kinds of errors: good email identified as spam, and spam identified as good emails. We need a filter that reduces both errors simultaneously.

Madigan shows that adding the word "Mortgage" as an indicator of spam will reduce the error rate: acceptance error is reduced from 240/635 = 38% to 140/485= 29%, and rejection error from 5/365 = 1.4% to 1/81 = 1.2%.

The suggestion is to use lots of words – "bag of words". In fact use all the words in the email! Of course you have to train the system with tables like Table 1 for "free" but for hundreds of words. However, if we progress with these words by constructing tables like Table 2 for many word combinations (i.e all combinations of hundreds of words) we quickly run out of patience – there are too many combinations. The "naïve" idea is to only use the info from the 2 x 2 tables for each word.

The simple example in the text of the naïve approach shows that the estimated odds ratio for identifying spam is 58 whereas the true odds ratio is 70. Don't worry too much about ,the details except that you should know what an odds ratio is. If an email has a probability of being spam of .8 and the probability of being a good email is .2, the odds ratio is .8/.2 = 4.

Summary: The article shows that the simple presence of certain combinations of words in an email can identify it as spam, and that the determination of these combinations can be done efficiently using 2 x 2 tables for the words in the email, and the odds ratio for each word. Modern computation makes the process fast enough to use it for all incoming email. The filter has error rates that are acceptable to most users.

Brookmeyer 197-210: Modeling an Outbreak of Anthrax.

This article shows that thinking statistically can assess the unrealized cost of a potential disaster. The funds spent to try to reduce the mortality from the disaster need to be justified, so that future occurrences can be handled efficiently.

The data set in this case is tiny – see Figure 3. In each of three exposures to the anthrax virus, the exposure was only identified after some cases were observed, and antibiotics were given to all those potentially exposed (about 10,000 individuals). The issue here is that the only latent times that are observed are the ones that occur before the antibiotics were served, and so were unusually short latent times. To estimate the number of individuals who were "saved" by the antibiotics requires an estimate of how many would have become ill if the antibiotics were not distributed.

The link between all the data in Figure 3 is that the exposure in each outbreak was assumed to be due to the same anthrax spores – the disease is the same. Of course each individual will react differently but we have no way of describing these differences except by postulating a probability distribution for it. The "parameters" of the distribution have to be estimated from the small data set. Maximum Likelihood is a way of estimating parameters.

Here is a very brief description of "likelihood" – the jargon word in statistics. Suppose a probability density, like the normal curve, has a formula f(x;p) which is a different function of x for each value of p. Now think of l(p;x) = f(x;p) as a function of p. This is the likelihood function of p. p is a parameter of the density of x. f(x;p), the density of x, describes the probability of x for a given p: the normal curve, for example, is higher in the center because more values occur there. But l(p;x) is not this curve – it is the curve with p on the x-axis, and it will be different for each observed x. The maximum likelihood idea is that the best value of p is the one that makes the observed x most likely. The thing to appreciate is that this approach depends heavily on the particular functional form chosen for f(). If this is justified, it is a powerful method.

Summary: This article shows that probability modeling can produce useful information even with a tiny data set. The assumption of a parametric model is quite a strong statement of prior information, but for decision-making, it is better than not doing any analysis.

(continued next page …)

De Veaux & Edelstein 307-322: Reducing Junk Mail Using Data MiningTechniques.

Data Mining: Extracting useful information from a large data base that has been collected ostensibly for some other purpose.

e.g. transactional – credit card
personal – membership registration
demographic – commercial supplier

Linkage is a big problem.

PVA – Paralyzed Veterans of America

Direct mail solicitations – address labels and greeting cards - cost per addressee = $0.68
Want to target addressees efficiently.

Model: response variables ; response?
amount of contribution?
And 481 potential predictor variables.
Suggests two models needed.
One at a time does not work well . Same for two-at-a-time.

Many strategies: trial and error is possible with computers! Important to use context.
Note that to test a strategy, need fit portion and test portion of data.

Summary: Data Mining has the potential to produce valuable information from data even when the data has been collected for some other purpose. However, the article shows that knowledge of the context of the data is essential for producing good information – it is not simply a matter of feeding the data to a large computer.

KLW 2010/02/09