STAT 100 **Notes** **February 11, 2010**

Today:

Sampling Error in public opinion polls
PVA Direct Mail article
Review of Assignment #4 (sample midterm I) and Midterm I

This morning's Vancouver Sun: 27% of BC residents are "very grieved" with Federal Government – 14% for Ontario. "Results accurate to within 2.5% 19 times out of 20." Based on a sample of size 1500.

SD of a proportion p in a random sample of size n = square root of (true p * (1-truep)/n)

Usual notation:

SD of $\hat{p} = \sqrt{p(1-p)/n}$

To be safe, use p=0.5 which makes SD largest. Then SD of $\hat{p}$ = √.5*.5/√1500 = 0.0129

So if we think of $\hat{p}$ as having a Normal distribution with mean p and SD 0.0129, then 95% of the time $\hat{p}$ should be within 2*0.129 of the true p. So true p = .27 ± .026 which could be described as within 2.6% 19 times out of 20!

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 (predictors) does not work well . Same for two-at-a-time (predictors).

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.

STAT 100 **Assignment #4 (Sample Midterm) Answers**

1. (8 marks) In the sports leagues we examined, there was doubt that the top team in the league was really better than the bottom team. How did we try to resolve this doubt?
A1. We simulated the league for the same number of games as was actually played, by assuming that every game was a 50-50 game. Then we compared the spread of points in the simulation with the actual spread in real life. The top team from the actual league had more points than the simulation would explain, showing that the top team really was better than the others.

2. (7 marks) The so-called "bell curve" or "Normal curve" does have the shape of a bell. How does this shape relate to the mean and SD of the normal distribution?
A2. Mean is in centre of bell, and SD is distance from centre to point of inflection (place where the curve down turns into a curve up).

3. (7 marks) What does Table 1 on p 364 suggest about alternative methods for short-term forecasting?
A3. It suggests that autoregressive method is best for this, since MAPE is smallest (see last para on p 364).

4. (8 marks) With reference to the article "Statistics in the Courtroom", why did the p-value of less than 1/1,000,000 suggest that Gilbert was guilty?
A4. The calculation of the p-value was based on the assumption that random error was the only source of the variability in the death-rate per shift. Since this was very small, and since the data was actually observed, we doubt the calculation, and conclude that our assumption was wrong, so there must be some other explanation. The association with Gilbert was one possible explanation.

5. (7 marks) In the Turkey mail article (Advertising as an Engineering Science"), how did the design try to reduce the likely effect of age of the email recipients?
A5. Age was treated as a covariate, and was adjusted for statistically.

6. (8 marks) Explain why averaging can make a portfolio of risky investments into a relatively stable portfolio investment.
A6. The average of the portfolio will tend to be positive as long as the long run average is positive, and the companies' returns are independent enough. It is the

square root law for the variability of means that ensures this is true as long as the portfolio is large enough.

7. (8 marks) Using Table 2 on p 420, how did the author of the article decide that investors could benefit from using regression analysis?
A7. The predicted relative prices (bottom line of table) are closer to the 1989 relative prices than the relative prices from the early 1970s. (Note that the predicted prices were specified in the harvest year, whereas by the early 1970s, experts already had several years to assess the wines. Still the predicted prices were the better guide for the investor.)

8. (7 marks) What outcome of a symmetric random walk (that is, the kind generated by a fair coin) is surprising to naïve observers?
A8. The surprising thing is that long runs up or down seem to quite likely to occur. (This suggests that the run pattern might be used usually for prediction.)

KLW 2010/02/02

STAT 100 **Midterm I** Feb 2, 2010

1. (8 marks) In the article "Randomness in the Stock Market", it is reported that "A set of randomly chosen stocks typically equals or outperforms the advice of the majority of investment newsletters". This is true even though most newsletter authors are highly-educated and respected in the investment business. Why do think this is so? (Hint: Think of a reason mentioned often in this course.)

A1. The symmetric random walk simulation showed that spurious trends seem to result even when there is no real predictability of the time series. It is likely that many people believe these trends are predictable, since it is counter-intuitive to realize this fact about time series that are close to symmetric random walks. (Even if the experts know this, the clients they serve likely do not).

2. (7 marks) Zipf's Law is an example of a model that is "wrong" but useful, when applied to the populations of a nation's cities. Explain.

A2. The model worked for Canada, and for the USA, but was very wrong for Australia. The useful outcome was that the urbanization of Australia was different from Canada and the USA, and so the analyses provide a starting position for a further look at the reason why.

3. (7 marks) In the article "Predicting Quality and Prices of Wines", why was it necessary to adjust for "AGE" before observing the influence of weather on the relative prices of vintage wines?

A3. The prices of wines will increase with AGE even if there were identical weather conditions in the years analyzed. But the weather variables do not have a

systematic increase with AGE. So to best see the relationship of price to weather, you need to look at the AGE-adjusted prices, in comparison with the weather data.

4. (8 marks) In the example of the five years of gasoline consumption that was used to illustrate the power of smoothing of a time series, it was mentioned that a moving average method could have been used for the smoothing. What determined the appropriate order of the moving average used for this purpose? Hint: The "order" of the moving average was the number of data values averaged at each step.

A4. The context of the data. The reasonability of the pattern provided by the moving average, noting for example that the annual-pattern might well be seasonal with period 12 months, is what determines how much smoothing is appropriate. Too little smoothing provides a chaotic pattern that is not credible, while too little smoothing decreases the amplitude of the seasonal pattern without changing the seasonal pattern otherwise.

5. (7 marks) In the Turkey mail article (Advertising as an Engineering Science"), why were no small p-values mentioned to support the claimed influences of SUBJECT and DAY-OF-WEEK on CLICK-THROUGH_RATE?

A5. The number of emails sent out is huge, so relationships noted in simple graphs are unlikely to be merely effects of randomness. However, there is reference to a logistic regression analysis in which proper hypothesis testing was done, but it was deemed to be too complex for the intended audience of the article (i.e. for beginning students of statistics). (This discussion on p 386.)

6. (8 marks) In your Assignment #4, question 6. asked you to explain "why averaging can make a portfolio of risky investments into a relatively stable portfolio investment". What are the requirements of a portfolio that make this a valid claim?

A6. A positive average return and independence of the investment outcomes of each company.

7. (8 marks) In the article "Advertising as an Engineering Science", there is frequent reference to the fact that the study described was an "experiment". What feature of the study made it a true experiment, **and** why was this study design chosen?

A7. The investigator assigned the features of the comparison groups to the experimental units – this is the feature (SUBJECT and DAY-OF-WEEK were assigned at random to the registrants.) This design was chosen because the investigator wanted unambiguous information about how these two variables affected (or "caused")the CLICK-THROUGH-RATE.

8. (7 marks) How would you use tosses of a fair coin to simulate equal-probability selection of the integers {1,2,3,4,5,6,7,8}? If you used this procedure to create a

sample of 25 values, estimate how likely would it be to get a sample mean less than 4.0? Hint: The SD of {1,2,3,4,5,6,7,8} is 2.5.

A8. Mean is obviously 4.5, and SD is given as 2.5. So 4.0 is 0.5 below the mean. The sample mean has SD = $2.5/\sqrt{25} = 0.5$. The normal distribution approximates the probabilities for the sample mean and the probability for the sample mean to be less than 4.0 is about 16%. (68% within 1 SD -> 32% outside of ± 1SD -> 16% below -1SD).

KLW 2010/01/29