STAT 100 Notes March 18, 2010

Today:

Midterm 2 Feedback
Survival Analysis: Aging of cars and people, traffic accidents.

Midterm 2 Feedback:

STAT 100 Midterm II March 16, 2010

Instructions: Answer all questions. Questions are approximately equal in mark value. You have 60 minutes for your answers – budget your time accordingly.

1. Three students each take a random sample of 25 pebble diameters from the gravel in a stream-bed. They report average diameters of 10mm, 11mm and 12mm. Estimate the SD of the pebble diameters in the stream-bed.
A1. The SD of the means is estimated as SD of {10,11,12}
$\sqrt{\frac{(10-11)^2+(11-11)^2+(12-11)^2}{3-1}}$ = 1 so the SD of the diameters in the stream bed is 5.
(SD of sample means = SD population /√n)

Further note: This question asks you to use the square root law in the opposite direction from the usual: SD of means estimates SD of pop, rather than the more usual SD of pop estimates SD of means.

Marker's Comment: Many students missed this one since they could not distinguish between the population standard deviation and the standard deviation of the means.

2. A large insurance company can drive a small insurance company out of business by reducing the premium for the identical policy. With reference to the theory of sample means, explain why this is so.
A2. The more policies an insurance company has, the less variable its average gross profit (average of (premiums-claims)). This is a result of the square root law for the variability of averages. By reducing premiums so profit is, on average, very small, the chance of a negative profit for a smaller company will be much larger than for a larger company. Negative profits for a few years will ruin any company.

Further note: As the sample size gets larger, the variability of means gets smaller. Applied here, as the number of policies gets larger, the variability of profit gets smaller, and when the average profit is positive, this means the chance of a negative profit gets smaller.

Marker's Comment: Many students did not answer the question posed.

3. Design advice in the article on brain response to memory tasks (pp 211-226) includes the statement "Use repetition, control what you can, and randomize the rest." Give one good reason for each of the three parts of this advice. You may use the specific context of the memory article if you find this helpful.
A3. Repetition allows for averaging which reduces variability of estimates.

Control the conditions of measurement, such as reducing the motion of the subject in the MRI machine, to reduce the unintended variability due to measurement error.

The effect of order of memory load treatments within a subject was unknown, and not of primary interest, so to balance its effect on the influence of the various memory load treatments on brain activity, the order was randomized.

Further Note: *Repetition* was achieved by having ten subjects (not just one). *Control* was achieved by trying to eliminate sources of error as much as possible, such as by strict instructions to subjects to be still in the MRI during measurement. *Randomization* was achieved by random orders of the various levels of memory load given to subjects.

Marker's Comment: The explanations of some students were not clear enough to reveal their understanding.

4. a) In designing a SPAM filter, what choice must a user select for the most satisfactory performance?
b) What practical problem does Madigan overcome in the design of the SPAM filter?
A4. a) the relative importance of the two kinds of errors: accepting spam as regular email, and rejecting regular email as spam.

b) The use of many words suggests considering which combinations are associated with spam. But Madigan shows that the information provided by one-word-at-a-time can be used jointly as an effective spam indicator.

Further Note: A very common feature of any decision-making situation is that there are two kinds of errors that must be balanced. In general they may be called "false positives" and "false negatives". Usually making one smaller makes the other larger. Madigan shows that two words are better than one word, but that the extension of this initial approach to many words becomes too complicated even for modern computers. So he proposes a different way to use many words that he calls Naïve Bayes – the essential feature of it is that the summary of SPAM ID information from each word used can be considered separately for each word, and these summaries combined. And this simpler method works well, apparently.

Marker's Comment: Question was answered fairly well – main problem was using the answer to part b) for part a) and then having nothing more to say about part b).

5. a) In a national pre-election survey of political opinion, what would be the main difficulty for the survey company if a random sample were wanted for the survey?
b) A national political survey in Canada is usually based on fewer than 10,000 respondents. Would a similar survey in the United States need a larger number of respondents? Explain.

A5. a) The main difficulty is to have a list of all eligible voters and a way to contact the random sample of them.
b) The survey would not have to be larger since the impact of the population size depends on the ratio n/N = (sample size/population size) and with n <10000, this ratio would be tiny in both countries.

Further Note: Some students mentioned general difficulties with a political opinion poll, like non-response or confidentiality fears. This is relevant but the question has an "if" in it that is important. It asks you to think about how a random sample of voters would be selected in a national survey. But this is almost impossible since a list of all voters rarely exists at one time prior to an election. And even if there were a list, contacting people from across the country at one time would be difficult.

Marker's Comment: Students mentioned general opinion poll survey difficulties rather than the one directed by the question.

6. In the article "Evaluating School Choice Programs" by Hill (pp 69-87), there was an opportunity to determine the relative merit of private vs public schools in teaching basic academic skills to elementary school children. While an experiment usually has a clear result in terms of the assessment of a causal factor (like type of school) on outcome (like performance in basic academics), some practical problems arose to make the inference less clear. What were these problems? (The general categories are all that is needed here, not the details).
A6. As is stated in the Conclusion on p 84, "missing data, noncompliance, and the sensitivity of results to choices made by the program evaluators." (Note to marker: OK in this instance to quote text.)

Further Note: This was mentioned in the review prior to the test.

Marker's Comment: This question was well done – the students who did not get full marks usually did not mention all three problems mentioned in the article.

7. What concept makes the "birthday paradox" seem not so surprising after all?
A7. When there are many chances for a rare event to occur, it is no longer a rare event when it happens. In the birthday paradox, there are many pairs of people in a small group of 25 persons – any one could be the match. (Note: think of the pairs in 5 persons: 12,13,14,15,23,24,25,34,35,45)

Further Note: The difference between this question and question 8 is that in this question, there are many chances for a single rafre event, while in question 8, there are many events, each of them rare, each with a small chance of occurring.

Marker's Note: Many students talked about p-values and hypothesis testing and did not answer the question posed.

8. Why do "coincidences" seem to occur so often?
A8. There are a huge number of events that, if they occur, would be considered "coincidences". If one of these occurs, it is reported as a coincidence. This is an example of "If a rare event has enough chances to occur, it will occur for sure".

Further Note: See Q 7.

Marker's Comment: Many students gave examples rather than an explanation.

New Material for week 9:

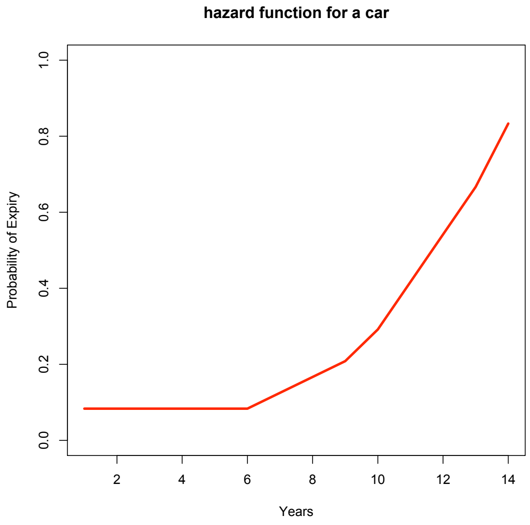
"9. **Survival**. Aging of cars, cells, and people. Traffic Accidents."

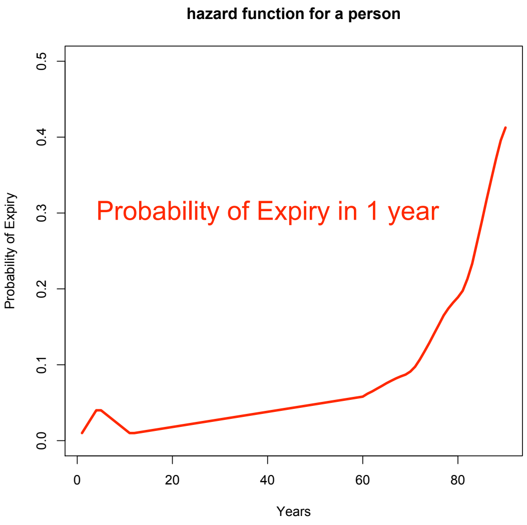
Survival Analysis

Wear in and wear out:

Think of the cars you are familiar with – how often did they need repair and how old were they at the time? Often new cars run without serious failings for a few years and then start to fall apart after the warranty expires! The cars "wear out". Next think of installing a new incandescent light bulb – have you ever had the experience of having it burn out either immediately or in a very few hours? If the bulb lasts the first few hours it usually lasts a long time, maybe 1000 hours or more. This situation is called "wear in" – its future life improves after a successful early period. Human infants are a bit like that too, although later in life they have a wear out phase too.

Some items seem to have a constant *hazard rate*: the rate per unit time at which they expire. Some electronic components have this feature – if I have such a component installed in 1000 different computers, the number that expire each month is approximately constant over, say, five years – maybe about 10 per month. The idea of constant hazard rate is clearly an approximation in this example, since eventually there will be less than 10 left to expire. Nevertheless, for predictions over the five year period, the assumption of constancy of the hazard rate is a pretty good approximation.





What does the hazard rate for a car, as a function of time, look like? If "expiry" means "off to the junkyard", it might look like the graph at left above. The graph on the right is for a person. Note that in both cases, the constant hazard function assumption would only be a reasonable approximation for a few years.

Note that probability of expiry is the complement of probability of survival. "Expiry analysis" is usually called "survival analysis". Of course, we need not have "expiry" as the end result of a "life". Here is a different scenario that is equivalent from the analysis point-of-view:

Suppose we are considering the "lifetime" as the time from getting a driver's license until the time that the driver is involved in a vehicle accident. "Birth" is the obtaining of the driver's license, and "expiry" is the occurrence of the first accident. "Birth" occurs at time 0. "Survival" at time t then means that, t months after the drivers license has been obtained, an automobile accident has not been experienced yet.

What sort of data do we need to estimate the monthly hazard involved in having a driver's license? To make things simple, let's assume that the hazard rate is constant over the time periods involved for students in the class – this is probably a reasonable assumption. In this case, the data we need is, for each student,

1) The month and year of your first driver's license issue
2) The answer to "Have you been involved in an auto accident?"

In 2) the intention is that any accident in which you were in a car, not necessarily as the driver, should count, including any that are not your fault.

For example, my answer to these questions would be

1) Sept, 1958
2) Yes

If everyone in the class today could write their answer to the two questions on a scrap of paper, and hand it in to me, I will do the analysis and circulate it by email, and put it on the course web page. Note that there is no ID required – just the two pieces of info above.

What will I do with the info on the scraps of paper?

Noting that I know today's date – March 2010 – I can calculate how long you have been a driver (we will ignore the potentially useful info about how much use of a car you may have had during the period). And the second part of the info tells me whether the "expiry" event – "accident involvement" in this case – has occurred. So I look at the proportion of students who have been involved in an accident among students with similar exposure times (grouped by 6 months say). Then I smooth this relationship (between *exposure time* on the x-axis and *proportion accidents* on the y-axis. From this graph I can infer the chance that a student in this class will be involved in an accident in the next month.

(In 2002 it was about 1%).

I hope this works out better than the randomized response survey!

Is it clear why this analysis is called "*Survival Analysis*"? A student survives an exposure of t months if the student has not been in an accident during that time. "Birth" is getting a driver's license, "Expiry" is being involved in an accident, and the survival time is the duration between Birth and Expiry.

Note that no single piece of paper will give me a survival time. All we know from a single student's info is whether the survival time is larger or smaller than the exposure duration. Only with the aggregation of all the info can we infer the distribution of survival times.

The most common applications of this kind of analysis are in the insurance industry and in clinical research. As an example of the latter, survival analysis is used in studies of which drug gives cancer patients the best survival distribution, or which surgical procedure does?

Assignment for week 9: No assignment!

I suggest you review the midterm carefully making use of the notes included here and asking yourself what you would have had to do to obtain more marks on the test. The final exam will be like Midterm 1 and Midterm 2, and it counts an important 50% of the course mark. Note that with a final exam that has an average numerical grade of about 50%, there is a lot of opportunity to raise your letter grade just from this one assessment.

KLW 2010/03/18