STAT 100 Chance and Data Analysis October 30, 2002

Today: Comments about Assignment 5 questions
More clarifications from survey
Assignment 6 (See last entry in these notes)

Comments about Assignment 5 questions:

Assignment 5: (Due 4:30 pm in correct box outside K 9510, October 30, 2002)

1. Use the Google search engine to find www references to Zipf's Law, and choose one source to demonstrate this "Law". Write up your findings in your own words in 100 words or less. One graph is probably needed as well – download it if you can, or submit a hand-drawn facsimile if you run into technical problems. Remember to reference the web site that you use (give the URL, the web address).

Important elements of your findings: graph of data, verbal description of result with reference to Zipf's Law, brevity!

2. In 100 words or less, explain the usefulness of random sampling in the context of one of the articles listed below:
CPI: pp 198-207
Census: pp 208-217
Jury: pp 87-92

Important elements: say how random sampling is used in the context chosen, and why it was useful in that context.

3. Consider the following lottery: Tickets numbering 1 to 1000 are sold for $1.00 each. 100 of the tickets return a prize worth $10.00. You buy 10 tickets. Describe the distribution of the prizes that your $10 worth of tickets will get you.

Important elements: Describe the distribution in any way you know how. (I told you you do not have to know how to compute the binomial probabilities.)

Back to Clarifications from Survey

2. Averages and Proportions – connections, mean & SD * * * * * * **

Theory about variability of sample averages (i.e. square root law) applies to sample proportions.
Why? Because a sample proportion IS a sample average. Just code the things you want the proportion of as 1 and everything else as 0. Proportion of 1s = avg of 0s and 1s.
See notes Oct 23 for details.
Note the useful shortcut for computing the standard deviation of a set of numbes that consists only of 0s and 1s. If p is the proportion of 1s in the set, then $\sqrt{p(1-p)}$ is the standard deviation of the numbers in the set. The way this is usually used is in estimating the variability of a sample proportion. Suppose a population has an unknown proportion p of items of a certain kind. We choose to code this kind as 1 and the rest as 0. Then take a random sample of size n from the population and call the proportion of 1s in the sample as p'. Then p is estimated to be p' and the variability of this estimate is estimated to have a SD of $\sqrt{p'(1-p')}/\sqrt{n}$. The reason for this is just the square root law for the variability of averages. Since p' is just an average of n things ….

For example, if I sample the SFU student population of 15000 using a sample size of 100 to record whether they have debt incurred to support thremselves while attending SFU, and suppose the proportion in the sample of 100 is 0.65. Then the proportion in the SFU popoulation is also esxtimated to be 0.65 but now we also know that this estimate is ± $\sqrt{.65(1-.65)}/\sqrt{100}$ = .048 or approx .05. We say the population proportion is .65 ±.05.

Randomized Response Technique * *(neg ests) *(calc) * *

81 students toss fair coin,
estimate that 40.5 get heads, and 40.5 get tails. (81/2=40.5)
So approx 40.5 answer Yes to question about coin outcome, and 40.5 answer question about marijuana use. We counted 46 Yes answers. So approx 46-40.5 = 5.5 were answering Yes to the sensitive question (about marijuana). But that was 5.5 out of approx 40.5, or 100 x 5.5/40.5 = approx 14%. So we ESTIMATE that the proportion of the class using marijuana weekly is about 14%.

(But note that the variability of this estimate will be mor complicated than for the student indebtedness example, since the estimate has an additional saource of variability – not

only from the sampling of the 40.5 students but also from the estimate of the proportion of the students that were answering the sensitive question.)

Q: What if we had less than 40.5 Yes answers? Instead of reporting a negative estimate for the proportion, just use 0.

4. Square root law – which n? * * * *

SD of avg = SD of things averaged / $\sqrt{number - of - things - averaged}$

5. Normal Model and Calculations * * *

We did this last time but I want to show you how the normal theory result held up pretty well ….

For a normal distribution,
Mean ± 1SD is typical range containing 68%,
Mean ± 2SD is 95%,
Mean ± 3SD is 100% (almost)
Remember, averages tend to have Normal Distributions.

Demo: Use deck of cards: shuffle and deal 5 to self. Record and pass on to next student. Use this coding to compute average: 2-10=2-10, J Q K A = 15 so the distribution of codes in the deck is

Code	Freq
2	4
3	4
4	4
5	4
6	4
7	4
8	4
9	4
10	4
15	16
---	----
all	52

Certainly not normal!
However the averages of 5 cards will be approx normal.

What about Mean and SD? If we assume we know the populalation
Can compute mean = 8.0 SD=4.7. So SD of average of 5 is $4.7/\sqrt{5}$ or about 2.1.
Compare with class result. (95% within 2.1 of 8, that is in the range 5.9 to 10.1?)

The rounded averages from the 5-card hands (with the 2-10 and 15 coding) were 4,5,6,7,8,9,10,11,12,13 with frequencies 2,2,4,6,5,4,6,4,3,3 respectively. Mean can be computed as 8.7 and SD=2.84 so mean ± SD is 5.9 to 11.5 and the proportion in this range is (4+6+5+4+6+4)/39 = 25/39 = 0.64. In other words 64% are within 1 SD. 68% was predicted by Normal theory. (Mean ± 2 SD in this case includes all 100 % of sample whereas theory predicts 95%). Note that normal theory is not expected to be perfect in this application for several reasons: sample size of 5 is small, random variation still plays a part, and rounding averages will distort outcome.

6. Binomial Model * * *

Random sampling of a sample of size n from of a a 0-1 population produces a sample with a number of 1s predicted by the binomial probability distribution. (We assume sampling with replacement, or sampling without replacement when n/N is small).
If we concentrate on the proportion of 1s, instead of the number of 1s in the sample, then the prediction can be approximated by the normal distribution. We use a mean and SD from the population (if we know it) or from the sample (if we don't know the population) to describe the normal population.

7. Sampling * * *
   Focus of sampling articles (Jury, Train Accounts) * * *
   Sampling with and without replacement

Focus is SAMPLING! (Hint: See course outline)

CPI: pp 198-207
Would it be possible to get the information necessary to construct an index of typical consumer prices without sampling products, retail companies, stores, days-of-the-week, etc?

Census: pp 208-217
In a population census, an attempt is made to count (and obtain some information from) ALL individuals in the population – so this is not sampling. But sampling IS used – read the article to find out how and why.

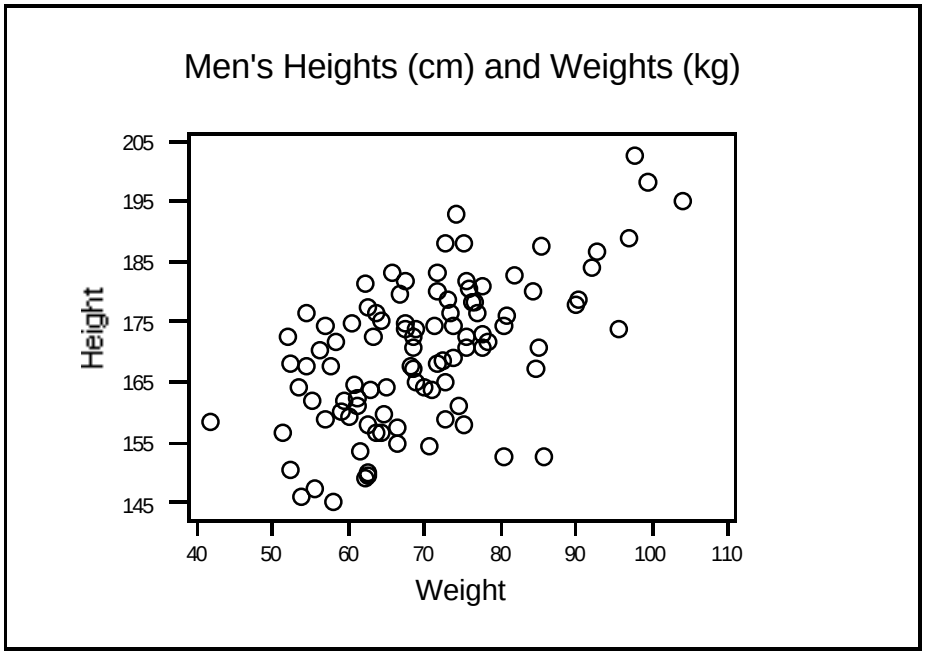
Jury: pp 87-92
The defendant's lawyer wanted to know what demographic characteristics would dispose a person to be prejudiced against his client. How can you find this out?

8. Correlation * *

This index measures the extent to which two variables are on the same side of their respective means.

i.e. Consider height and weight. My weight is 100 kg. Weights for males average about 70 kg and SD would be about 12 kg, so I am about 2.5 SDs above average. If men's height and weight were perfectly correlated (correlation coefficient = +1), I would be 2.5 SDs above average in height. If men's heights averaged 170 cm and SD about 12 cm, then I would be 200 cm (2 m) tall!

But men's heights and weights are not perfectly correlated, and I am only a tiny bit taller than average, unfortunately. Here is a picture of data determined by my assumed distributions and a correlation of .6.



9. Choice of study designs *

The big choice is experimental vs observational

experimental is better for proving causal links, observational is cheaper
Often observational is the only one feasible for reasons other than cost.

10. Solar System article *

Just understand the way in which the correlation coefficient is used here. Not necessary to master the whole article.

Assignment 6 (Due Wed., November 6, 4:30 pm in boxes outside AQ 9510)

1. Suppose a certain model of car built in 1985 survives for an average of 10 years – that is after some time it arrives at the junk yard at various ages which average 10.0. Of the 200 such vehicles sold that year, most have by now "died" but a few are still going.
a) Draw a graph (by hand or by computer) that you think would approximate the histogram of the age-at-death of the 200 vehicles. Use relative frequency to scale the height of the histogram rectangles.
b) Using your graph from part a), estimate the number of vehicles that last more than 5 years, more than 10 years, and more than 15 years.
c) Of the cars that last for 5 years, what proportion of them last an additional 5 years?
d) Of the cars that last for 10 years, what proportion of them last an additional 5 years?
e) Comparing c) and d), say something about the aging rate of cars implied by your guess in part a)?

2. In an application of the randomized response technique, a sensitive question might be "Do you currently pay any income tax in Canada?" Suppose that in a class of 500 students, each student rolls a die to determine if they will answer Q1 or Q2 below:

Q1. Did you get a 1 or a 2 face-up on the die?
Q2. Do you currently pay any income tax in Canada?

If the student gets a 1 or a 2 on the die, they answer question 2. If they get a 3,4,or 5, they answer Q1. If they get a 6, they roll again. The instructor asks for the Yes answers to the selected questions and counts 400 "Yes" responses.

a) Estimate the proportion of the 500 students that would have answered Yes to Q2.
b) If a oridinary survey of the class had selected a random sample of 200 students and obtained the proportion of Yes responses by guaranteeing anonymity in some other way (so you can assume truthful answers),what would be the SD of the estimated percentage responding Yes?
c) Do you think this SD in b) is larger, smaller, or about the same as the SD of the estimate in part a)? Say why.