STAT 100 Chance and Data Analysis October 23, 2002(rev)

Today: Randomized Response Technique
Averages and Proportions
Feedback Form
Lotteries (postponed to Oct 25)
Assignment 5 (not postponed!)

Randomized Response Technique:
A method for getting honest responses to a sensitive question …
Consider the two questions …
Q1: Did your coin toss come up heads?
Q2: Do you smoke marijuana at least once per week?
You toss a coin to see which question you will answer. If your coin comes up heads, answer Q1. If it comes up tails, answer Q2.

If you answer "Yes" or "No", nobody knows whether you are answering Q1 or Q2, as long as they do not see your coin toss outcome. This guarantees that your true answer to the sensitive question will be protected from the others in the class.

Note that the subset of people who do answer the sensitive question is a random sample of size about 50 from the class of 100 (because it is selected by toss of a coin).

If there were 100 students participating in the survey, and suppose 60 answer "Yes". Although we do not know exactly, approximately 100*(1/2)=50 will answer Yes because they got a Head on the coin. So we would guess that, of the other 50 who got tails on the coin toss, and answered Q2, 10 answered Yes to the sensitive question. So we estimate, 10/50 or 20% of the class smokes Marijuana at least once per week.

Any other sensitive questions you would like to try? (Suggestion from my daughter "Do you drink alcohol just to fit in?")

Randomized Response Technique is an example of use of randomness to provide confidentiality. A more practical version of this the addition of random noise to data files about individual people or companies so that researchers can have access to the data files without invading privacy of the people or companies. As long as the statistical properties are preserved, the data is still useful.

Now we return to a small technical question: how much variability is there in the proportion of the class that do get a head on the coin toss?

Think of the coin toss as a random sample of size 1 from the population {T,H}.
100 such tosses would result in a list of 100 responses such as H,H,T,H,T,T,H, ….,T.
For the purpose of counting Hs, let us re-code this sampling process as sampling from a population {0,1}, so our responses would be written instead as 1,1,0,1,0,0,1,….,0.

The technical question is now: What can we say about the proportion of 1s in a sample of 100 selected with replacement from the population {0,1}? (i.e. mean and SD=?)

We already have some theory about this because the proportion of 1s is exactly the same as the average of the sample of 0's and 1s. (A small scale version of this for the sequence 1,1,0,1 is that the proportion of 1s is 3/4 and the average is (1+1+0+1)/4=3/4 also.) But we know that averages have an SD = (individuals SD)/$\sqrt{n}$ where n is the number of things averaged.

In the application described here, the individuals SD is the same as the SD of {0,1} which is $\left([(0-(1/2))^2 + (1-(1/2))^2]/2\right)^{1/2}$ = 1/2 so the SD of the average, and of the proportion of 1s, is (1/2)/$\sqrt{100}$ = .05

This means that the proportion of heads in 100 tosses of a fair coin averages 0.5 but has an SD of .05. So we might write number of heads in 100 tosses = 50 ± 5.
This answers the question posed.

We could study how this affects our estimate of 20% above, and how variable that estimate might be, but this is not essential for this course.

Note: Here is a short cut for the SD of a set of 0-1 numbers. If the proportion of 1s is p, the SD is $\sqrt{p(1-p)}$. So for example the SD of {00011111) is $\sqrt{(5/8)(3/8)}$ = .48

If we do not know the proportion of 1's in the population sampled, we use the SD of the sample as an estimate. For example, if in a random sample of fifty newly admitted students to SFU 10 say they want to be Crim majors, we would estimate the proportion of Crim major aspirants among new admissions as 10/50 =0.2 but the ±SD would be estimated as ± $\sqrt{.2(.8)}\Big/\sqrt{50}$ = .06 approx so the estimate might be written as .2 ±.06

What is important so far?

1. Randomized Response Technique is an example of using randomness to provide confidentiality.
2. Sample proportions have the same statistical properties as sample averages.
3. The SD of sample proportions is easy to calculate if the population proportion is known, and easy to estimate if the population proportion is not known.
4. Note that an estimate is better if it has a variability indication attached (i.e ±SD)

Next: Lotteries:

Last 2 digits of your student number can be your lucky number.

Use Cards with A=1, 2-9, 10=0 and select winner of "twoney". If more than one winner, need to pay out more than one twoney.

What should I charge to run this lottery? How much for tickets?

Chance of winning is 1/100. Cost for break-even is 2¢ per ticket.

Does this lottery provide any money for community services?

Need to charge 4¢ per ticket to net 2¢ gain on average. More next time …

Feedback Survey:

1. Most Important Idea learned since last survey (Sept 27)
2. Most Confusing Topic presented since last survey
3. Topic I would like to have additional material about
4. Anything else you want to say ….

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).

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

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.