STAT 100 Chance and Data Analysis Oct. 21, 2002 (Rev)

Today: Sampling – general intro
Accounts article (pp 151-160)
Jury Selection (pp 87-92)
CPI (pp 198-207)
Census (pp 208-217)

A first look at “Sampling”:

The idea is to use examine part of something to tell you about the whole. Key is how to choose a “representative” part.

For example:

• samples of food products (esp. meat) are analyzed to test for contamination
• 2 by 4 s are sampled to test their breaking strength as part of a quality control process
• auditors sample accounts to check for errors (size and frequency)
• drivers are sampled for alcohol level
• whale populations are sampled (as in Whales article).

Group sampled = “population”
Selected subset = “sample”
Number selected is called the “sample size”
Method of selection = random sampling (usually, for this course)
Method of summary: dotplots, or means and SDs, usually.

Example: Population of digits in which each digit is equally represented, sample of size 7, summarize by average digit and SD of digits, or dotplot (as we did).

Usually samples are selected “without replacement”, from large populations. (Like census of population).

Note that sampling involves variation, and while the variation is easily explained as due to the sampling mechanism, it is uncontrolled, and so we must learn to live with uncontrolled variation just as we did with unexplained variation. Another type of “UV”.

Accounting article (pp 151-160)

Sampling accounts receivable

Chesapeake and Ohio Railroad Co.

Allocation of freight revenue to several railroad companies.
23,000 waybills (records that describe each shipment and charges)
total charge is known, but not the allocation to C&O. Need to examine the waybills to find out. Try sampling. Startified random sampling in this case.

stratum of size of charge on waybill
Different sampling rates depending on variability in strata.
Must correct for different rates to get total allocation.

$0-5	1%
$5-10	10%
$10-20	20%
$20-40	50%
$40+	100%

Simple random sampling within each stratum. Use of random numbers.
Use of serial numbers in this case.

2072 waybills were sampled out of 22984 total. (9% of them)

Suppose in $0-5 sample, there were 5000 waybills and 50 were sampled. And in those 50, 15% of the charges were allocated to C&O. So we know that a proportion (5000/22984) * 0.15 of the total charges of all waybills are owed to C&O from this category alone. We can do the same for each stratum to estimate the total owing to C&O. And we still have not looked at over 20,000 of the waybills.

The article shows that the method produced a very accurate estimate since in this case, a complete census was eventually done to check the procedure….

Similar situation for airlines and radio royalties. See article.

Some theory:

If I have a population of 5 numbers: 1,2,3,4,5

Suppose I take a random sample of size 2: 3,5 say

Will the mean of a sample of size two vary differently depending on whether I sample with replacement or without replacement? Yes. It will be less when sampling without replacement. To convince yourself of this, consider taking s ample of size 5!
Here is the general case:

SD of average in sampling without replacement: n sample size; N population size

SD of mean = SD of individual values/ (sqrt n) times $(1-(n-1)/(N-1))^{1/2}$

$(1-(n-1)/(N-1))^{1/2}$ is the "finite population correction factor"

Note that (n-1)/(N-1) is almost n/N. When this is close to 0, the correction factor is 1 (i.e correc tion factor of 1 means no correction)

For a sample size that is a small proportion of the population, ignore correction.

Idea: if the sample is small relative to population, chance of sampling twice in sampling with replacement is negligible, so sampling without replacement and sampling with replacement are essentially the same in this case.

The important thing is to realized the following:

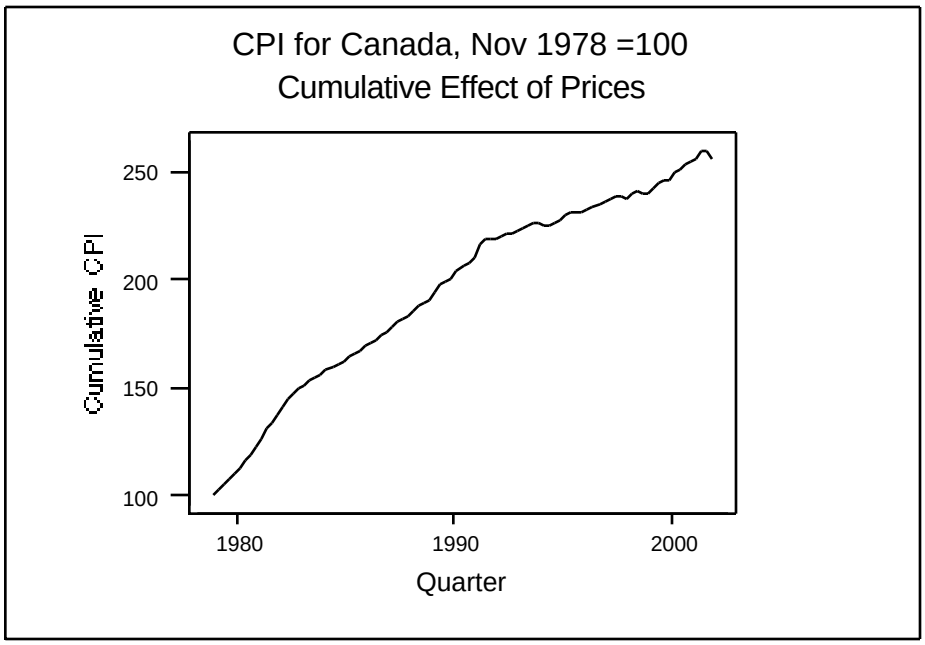
The precision of a sample mean (and also the SD of the sample mean) is determined mainly by the sample size, not by the population size. The exceptions to this do not arise in practice: small populations are not sampled, and samples are usually a small proportion of large populations.

Jury Selection: (pp 87-92)

Socio-demographic characteristics relate to attitude toward case – if these are known, defendant's lawyer can avoid prejudiced jurors. Get by sampling survey.

CPI (pp 198-217)

Sampling is the only way to get an idea of the general cost of things. What basket of goods will be used? What about infrequntly bought items like vehicles or houses? Will the basket change over time? Which prices will be used? In the revised notes I will include a graph of the CPI index in Canada for 1975-2001.



Note: Costs will double every n years, where n=70/i. So 7% per year will cause costs to double every 10 years.

Census (pp 208-217)

"Census" means 100% sampling. But our Canadian census does do sampling – usually about 5% of households are required to fill out a more detailed form. The cost is reduced to about 5% of what a census would caost for this more detailed data.

Possible Assignment or Midterm 2 question: Explain the use of sampling in the context described in the article on ….. More detail that I have given would be required on this open book midterm.

Assignment 5 will be assigned on Wednesday for submission Oct 30.