STAT 100 Chance and Data Analysis Nov 6, 2002

Today: Review for Midterm 2

Major topics since MT1:

Sampling wild animal populations - understanding sampling
Correlation coefficient – how closely two variables are related
Zipf's law – finding regularity in a chaotic world
Binomial probability distribution model – useful model for counts
Normal probability model – useful model for averages and proportions
Sampling in several contexts – understanding the usefulness of random sampling
Sampling with and without replacement – implications of the difference
Using 0-1 populations: variability of sample proportions (and square root law)
Randomized response technique – another sampling exercise
Lotteries – how they work - average net returns
Survival analysis – strategies for estimation from partial data

More detail: Important Methodological Techniques

1. Sampling Framework:

Population - what we want to know about E.G. SFU undergrad population
Sample - the part of the population that we have selected
Sampling Method - usually random so can predict variability
Numerical summary of Sample – often mean (or proportion) and SD, or sometimes a graph
Derived Information - often the variability of the mean (or proportion)
Purpose of mean and its variability? To describe the population mean.

Example: Population: Voters preferences for candidates in municipal elections Nov 16
Sample: Based on "Sampling Fram" - Voting lists?
Sampling Method: Possibly random sampling
Numerical summary: Mean and SD (Use 0-1 formula if appropriate)
Derived Information: proportion preferring candidate A in sample and its SD

2. Sampling Variability in 0-1 populations

Supposea large population is 55% type A and 45% other types (members of the Yuppee Club?)
Consider the proportion of type A in a sample of size 25 from the large population.

There are two situations we know how to deal with:

Situation 1: The 55% is unknown to the investigator but the sample proportion is observed, say 15/25 = .6.
The sample proportion of 0.6 is an estimate of the poulation proportion of type A.
i.e. we estimate 60% type A in the population. But what is the precision of this estimate?
An equivalent question is What is the variability of this estimate? (Important idea).
The variability of a sample proportion is estimated in this case by
(the estimated population SD) / $\sqrt{25}$
The estimated popoulation SD = $\sqrt{p(1-p)} = \sqrt{(.6)(.4)} = .50$
So the SD of the sample proportion is approx .50/5 = 0.1
The estimated population proportion is .6 ± .1

Situation 2: The 55% is known and so we are not trying to estimate it. But we may be interested in what the variability would be IF p=.55.

SD of sample proportion =
(the SD of the population) / $\sqrt{25} = \sqrt{p(1-p)} \Big/ \sqrt{25} = \sqrt{(.55)(.45)} \Big/ \sqrt{25} = .49/\sqrt{25} = .098$
and so the distribution of the sample proportion is described as having mean .55 and SD =.098 exactly.

This is not an estimate!

3. Distribution of means and proportions – Normal Distribution

Averages tend to have a Normal distribution (so do sums):

What do you have to know about Normal distributions?

Bell shaped (symmetrical, not skewed) 68% within 1 SD, 95% within 2 SDs, 99.7% within 3 SDs.

4. Correlation Coefficient

Closeness to a line of the points in a scatter diagram of two variables (e.g. height and weight)
Average product of variables after they have been "standardized" (subtract mean and divide by SD for each variable).
close to ±1 if strongly related (linearly) and to 0 if not.

Does not measure non-linear relationships.

Other Contextual Familiarity:

animal population estimates
Zipf's law applications
Uses of sampling in Jury, CPI, Accounting, Census, ...
Lottery ideas
Survival Durations (car accidents)