Lecture 34: A look Ahead

1. Conditioning
2. Symmetry
3. Random variables and their distributions
4. Stories

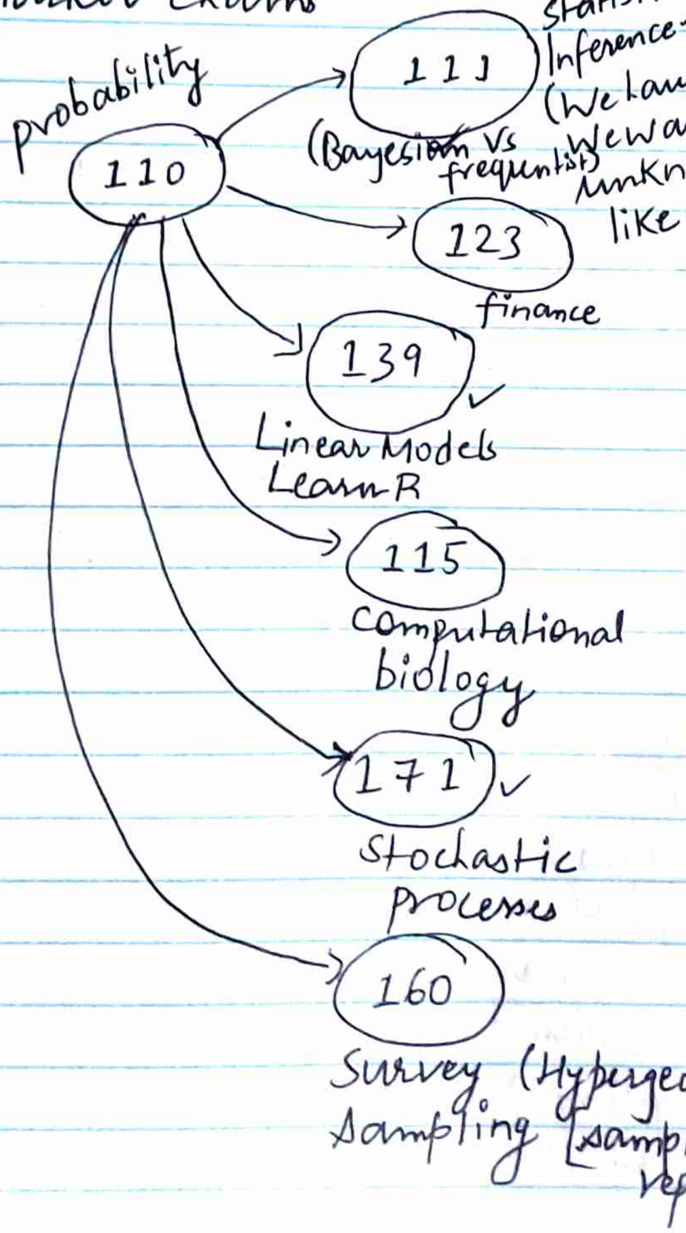
What is randomness and uncertainty?

5. Linearity
6. Indicator random variables
7. LOTUS

Computing Expected values/averages.

8. Law of Large Numbers (LLN)
9. Central Limit theorem (CLT)
10. Markov chains

Long run behaviour


probability
110
111
Statistical Inference
(Bayesian vs frequentist)
(We have data, then We want to estimate unknown parameters like θ or λ, etc.)
Like we have lots of observation so far, but then in the future we are going to get more data and we want to predict what's that going to look like.
123
finance
139 ✓
Linear Models
Learn R
115
Computational biology
171 ✓
Stochastic processes
160
Survey (Hypergeometric) Sampling [sample without replacement]

Books

Mostly Harmless Econometrics
- → Conditional Expectation
- → Adam's law
- → Eve's law

Sampling from finite population

Let's true values (of something that we are interested in, i.e. for each person we have some variable. It could be their height, their income, their opinion on some question, whatever we are studying) are $(y_1, y_2, \ldots, y_N)$, where N is the size of the population.

And, let's assume each person has an ID number like social ~~sec~~ Security number, so there's some well defined way to list them out.

→ treated as non-random, fixed (constant).

Sample of size n, goal is to estimate the average of $y_1, y_2, \ldots, y_N$, i.e., $\sum_{j=1}^{N} y_j$

$\to p_j > 0$

prob. that person j is in the sample is p_j (Known).
So, the simple random sampling would be the case when all the p_js are equal, i.e., everyone is equally likely to be collected into our sample (But, obviously this may not be true. some people may be much easier to sample than the others or some are just obscure, hard to reach, etc. So p_js may not be all equal in practice)

Let $(X_1, Z_1), \ldots, (X_n, Z_n)$ be the sample data;
X_j is the y value we are interested in. Xs are random variables and ys are fixed. The Xs are random because this is the first we collect into our survey, but that person was randomly chosen with some probabilities, so the value has become random because of the sampling;
Z is their ID number of that person, i.e. who did we actually get.

Unbiased estimator for the total

Then $\sum_{j=1}^{n} \frac{X_j}{P_{Z_j}}$ is unbiased.

(This says take each measurement and divide by the prob. that we actually got that person in our survey, then that's unbiased.)

Z_j is the random id number So, in the denominator we have random probability

Proof

$$\sum_{j=1}^{n} \frac{X_j}{P_{Z_j}}$$

$$= \sum_{j=1}^{N} I_j y_j / p_j$$

Over entire population

, where I_j is the indicator of jth person being included.

Expected value : $E\left(\sum_{j=1}^{N} \frac{\cancel{E} I_j Y_j}{P_j}\right)$

By linearity

$$= \sum_{j=1}^{N} \frac{E(I_j)\, Y_j}{P_j}$$

The expected value of I_j, by defn. and fundamental bridge is the prob. that person j is included in the sample. $= P_j$

$$= \sum_{j=1}^{N} \frac{\cancel{P_j}}{\cancel{P_j}} Y_j$$

$$= \sum_{j=1}^{N} y_j$$

Horwitz-Thomson Estimator or, Inverse probability waiting

Is itagood estimator?

EX Basu's Elephant

Sampling with Replacement and sampling without Replacement

When we sample with replacement, the two sample values are independent. Practically, this means that what we get on the first one doesn't affect what we get on the second.
Mathematically, this means that the covariance between the two is 0.

In sampling without replacement, the two sample values aren't independent. Practically, this means that what we got on the for the first one affects what we can get for the second one.
Mathematically, this means that the covariance between the two isn't 0. This complicates the computations.

When we sample without replacement, and get a non-zero covariance, the covariance depends on the population size. If the population is very large, this covariance is very close to 0. In that case, sampling with replacement isn't much different from sampling without replacement. In some discussions, people describe this difference as sampling from an infinite population (sampling with replacement) versus sampling from a finite population (without replacement).