1 Lecture 3 Monday 01/22/01

Homework 1:
chap. 7, exs : 1, 8, 19, 22, 45, 50, due next wednesday in class.

Sectionning: see the logictics section

3.1 Review of last time, if you missed see: Lecture 2

We studied the properties of the estimator $\bar{X}$ obtained by taking the average of s simple random sample.

It is unbiased, and it's variance is given by:

$$Var(\bar{X}) = \sigma^2(1 - \frac{n-1}{N-1})$$

This formula contains a quantity that is usally unknown σ^2, how are we going to estimate it?

3.2 Estimation of Population Variance

Sample survey is used to estimate the population mean for instance (it's a parameter), next step that makes sense is if we don't actually know either mean or variance, we have to estimate them. If we are estimating just the mean, we will need an estimate of the variance anyway in order to evaluate it because we only know the Var(mean) as a function of σ.

Population variance is the average squared deviation from the mean, the first estimator that comes to mind is the average squared difference from the sample mean for :

This is actually biased, it is our first example of a biased estimate,

Theorem 1 *With simple random sampling*

$$E(\hat{\sigma}^2) = \sigma^2(\frac{n-1}{n})(\frac{N}{N-1})$$

Proof

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\bar{X}^2)$$

Remember the general formula, for any rv W:

$$E(W^2) = var(W) + E(W)^2$$

Applying this to X_i and $\bar{X}$ gives

$$\begin{aligned} E(X_i^2) &= \sigma^2 + \mu^2 \\ E(\bar{X}^2) &= var(\bar{X}) + [E(\bar{X})]^2 \\ &= \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1}) + \mu^2 \end{aligned}$$

This gives :

$$\begin{aligned} E(\hat{\sigma}^2) &= \sigma^2 + \mu^2 - \mu^2 - \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1}) \\ &= \frac{N}{N-1}\frac{n-1}{n}\sigma^2 \end{aligned}$$

And as we always have $n << N$ we will have

$$\frac{N}{N-1}\frac{n-1}{n} < 1$$

The natural estimate for the variance underestimates it, why ?

Well the sample mean fits the sample better than it should because it is computed from the sample, this is a case when we have an opimistic evaluation because we use the same thing to estimate as we do to elauate, we will see later in cross validation techniques there are ways of avoiding this.

$\hat{\sigma}^2$ underestimates σ^2 on average.

Especially for small samples, the sample clings more closely to $\bar{X}$ than it would to μ.

We can correct this because we KNOW the bias (a rare circumstance), so we apply the inverse of the factor this will give us the corollary:

Corollary:

An unbiased estimate for $Var(\bar{X})$ is :

$$s^2_{\bar{X}} = \frac{\sigma^2}{n}\frac{n}{n-1}\frac{N-1}{N}\frac{N-n}{N-1} = \frac{s^2}{n}(1 - n/N)$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

This is called the unbiased estimate of the variance ingeneral in computer packages and calculators when you ask for the variance of a vector of numbers, they think this is a sample and use this formula, the 'unbiased estimate'.

3.3 The Normal Approximation of the Sampling Distribution of $\bar{X}$

We have estimates for the two first moments, what more could we say, well we'll see if we also knew the sampling distribution, that is the proababilty distribution of the mean, we could then say alot more.

If the sampling had been with replacement the mean would be a sum of iid variables, and we could happily say that when the sample size was big enough then the distribution of the sample mean would be close to that of a Normal variable.

It's important enough that I am happy to remind you of the central limit theorem at this stage:

Convergence in distribution:
Let $X_1, X_2, ...$ be a sequence of random variables with cumulative distribution functions $F_1, F_2...$ and X a rv with cdf F Then we say that the sequence X_n converges to X if

$$lim_{n\to\infty} F_n(x) = F(x)$$

every x where F is continuous.

CLT:
Let $X_1, X_2, \ldots$ be a sequence of independent r.v.'s with mean 0 and variance $\sigma^2$and the common distribution F (and extra techincality:moment generating function M defined in a neighborhood of 0).

Let $S_n = \sum_{i=1}^{n} X_i$ then

$$lim_{n\to\infty} P(\frac{S_n}{\sigma\sqrt{n}}) = \Phi(x), -\infty < x < \infty$$

Consequence for the mean :

$$lim_{n\to\infty}(\frac{\sqrt{n}\bar{X}}{\sigma} \leq x) = \Phi(x)$$

Theorem (without a proof here): In simple random sampling :

$$lim_{n\to\infty}(\frac{\sqrt{n}\bar{X}}{\sigma} \leq x) = \Phi(x)$$

————————————end of lecture 3————————————-