1 Lecture 4 Wednesday 01/24/01

Homework and Sectionning: see the logictics section

3.3 Review of last time, if you missed see: Lecture 3

We computed an unbiased estimate for $Var(\bar{X})$ as : (typo in lect 3 notes)

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

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

Then we went further towards complete knowledge of $\bar{X}$'s properties, its complete distribution:

3.4 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 proabability 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 :

$$lim_{n\to\infty}(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \leq x) = \Phi(x)$$

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

$$lim_{n\to\infty}(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \leq x) = \Phi(x)$$

What is this useful for ?

If we want to compute the probability that $\bar{X}$ is within a certain distance δ from μ.

$$\begin{aligned} P(|\bar{X}-\mu| \leq \delta) &= P(-\delta \leq \bar{X}-\mu \leq \delta) \\ &= P(-\frac{\delta}{\sigma_{\bar{X}}} \leq \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \leq \frac{\delta}{\sigma_{\bar{X}}} \\ &= \Phi(\frac{\delta}{\sigma_{\bar{X}}}) - \Phi(-\frac{\delta}{\sigma_{\bar{X}}}) \\ &= 2\Phi(\frac{\delta}{\sigma_{\bar{X}}}) - 1 \end{aligned}$$

because $\Phi(-z) = 1 - \Phi(z)$

Example :
$N = 393$ hospitals, take n=64, We can compute $\sigma_{\bar{X}} = 67.5$.

What is the probability that the population mean is off the sample mean by more than 100 either way ?

$$\begin{aligned} P(|\bar{X}-\mu| > 100) &= 2P(\bar{X}-\mu > 100) \\ P(\bar{X}-\mu > 100) &= 1 - P((\bar{X}-\mu) < 100) \\ &= 1 - P(\frac{(\bar{X}-\mu)}{\sigma_{\bar{X}}} < \frac{100}{\sigma_{\bar{X}}}) \\ \approx 1 - \Phi(\frac{100}{67.5}) &= .069 \end{aligned}$$

So that the probability is around .14 that it's off by more than 100 either way.

This gives us a statement about an interval with which we associate a confidence level.

Definition: A confidence interval for a parameter θ is a random interval calculated from the sample that contains θ with some specified probability.

In our case study we knew μ, we don't usually, the distributional study allows us to make such statements as:

$\bar{X}$ is within 1.96 SE 's of μ, 95 % of the time, which can be written :

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{s_{\bar{X}}} \leq z_{\alpha/2}) \doteq 1 - \alpha$$

Now at this point there is a trick in that we turn the intervals inside out to give the probability of catching μ.

The probability that μ lies in

$$\bar{X} \pm z_{1-\alpha/2}\sigma_{\bar{X}}$$

is approximately $1 - \alpha$.

Notation reminder :

$z_{\alpha/2}$ is the $\alpha/2$th quantile of the Normal (0,1) distribution. it is the inverse of the cdf taken at $\alpha/2$, otherwise known as $\Phi^{-1}(\frac{\alpha}{2})$.

Here is an image that may help the understanding of this trick for the classical confidence interval reasonning:

Suppose we have an archer whose precision we know in the following way, she hits the 10cm bull's eye 95 % of the time. (1 miss in 20)

We were standing behind the target the arrow is shot, how would we choose to estimate the position of the bull's eye ?

If for each shot a circle of radius were drawn, the circle would include the center 95 % of the time.

So we could draw such a circle around her shot and we could say we were 95% sure of having the center in there.

It's a two-step procedure: alot of shots
probability calculations give the width of the bull's eye.

The arrows are images for the estimates θ. The center is μ.

The intervals are random because the estimates are, and as they are the centers of the intervals...

Actually even more is random σ^2 being unkown the actual width of ci is random too. (For large samples this doesn't have any effect).

How large is large ?
It's around $n = 20$

Usually α is of an order of the percent, or 5%. We only want to miss μ rarely.

Example :

We could check this theory by making random samples from our hospital data : (p. 204) They actually all covered μ.

Let's try with a matlab program:

```
function cis=ci(datav,n,S,alpha)
%Function that estimates the mean
%by drawing a sample of size 16
```

```
%from the population defined in data
%then computes the se(mean)
%and S  (1-alpha) confidence intervals
%then plots them
%Find the number of observations in the population
Npop= length(datav);
%Take a random sample of size n from data
cis=zeros(S,2);
for s=1:S
    rp=randperm(Npop);
    ri=rp(1:n);
    sample=datav(ri);
    xbar=mean(sample);
    sd=std(sample);
    stderrm=(sd/sqrt(n))*sqrt(1-n/Npop);
    zalpha2=norminv(1-alpha/2);
    cis(s,1)=xbar-zalpha2*stderrm;
    cis(s,2)=xbar+zalpha2*stderrm;
end
    plot(cis,'o');
    for s=1:S
              line([s,s],cis(s,:));
    end
```

Note: To use the program, save all the above lines in a file in the directory from which you run matlab as **ci.m**, then try it by the lines

```
load('hosp.dat');
hosp1=hosp(:,1);
hosp1=hosp1';
c1=ci(hosp1,30,100,.05);
```

The width of a confidence interval is determined by the sample size, the population standard deviation and the confidence level chosen.

We are going to see some examples. Apart from knowing some of the quantiles of the standard normal by heart, you can look them up in the tables at the back of the book, you can also use the following matlab command:
norminv(.005)