Math 152 1/27/09

Simple Random Sampling (Cont.).

Last time we introduced the notion of biased and unbiased estimates and saw that

$$E(\bar{X}) = \mu$$
$$E(T) = \tau$$

so that $\bar{X}$ and T are unbiased estimates of μ and τ respectively.

Recall also that on the way to the above we saw that

$$E(X_i) = \mu$$

and $Var(X_i) = \sigma^2$.

μ = true population mean

σ^2 = true population variance.

What is $\mathrm{Var}(\bar{X})$?

$$= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right).$$

If the X_i were independent (e.g. if we sampled with replacement) this would be

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}.$$

and we would have, for the standard deviation of $\bar{X}$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

This <u>almost</u> true. The X_i, X_j $i \neq j$ are only very weakly dependent if n is small compared to N.

$$Var(\bar{X}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2} Cov\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j)$$

By the calculation made in the proof of Lemma B on pg 207 of your text

$$Cov(X_i, X_j) = -\sigma^2/N-1$$

if $i \neq j$.

(See problems 25 and 26).

So

$$Var(\bar{X}) = \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j}\sum Cov(X_i, X_j)\right)$$

$$= \frac{1}{n^2}\left(n \cdot \sigma^2 + (n^2 - n)\left(\frac{-\sigma^2}{N-1}\right)\right)$$

$$= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}} .$$

$$\simeq \frac{\sigma}{\sqrt{n}} \quad \text{when} \quad \frac{n}{N} \text{ is small.}$$

<u>e.g.</u> hospitals (discharges).

$n = 32$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}$$

$$= \frac{589.7}{\sqrt{32}} \sqrt{1 - \frac{31}{392}}$$

$$\simeq (104.2)(.96).$$

$$\simeq 100.$$

On the figure (produced in class using R or in your text) most of the observations (out of 500) were within two standard errors of the mean. (814)

i.e. in (614, 1014).

e.g. estimating proportions

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{1-\frac{n-1}{N-1}}$$

with $n = 32$, $N = 393$.

$$\sigma_{\hat{p}} = \sqrt{\frac{.654 \times .346}{32}}\sqrt{1-\frac{31}{392}}$$

$$\simeq .08.$$

Compare with the computer simulation.

Though $\sigma_{\bar{X}} \simeq \frac{\sigma}{\sqrt{n}}$ is nearly independent of N, we aren't as fortunate for σ_T

$$T = N\bar{X}$$

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

In applications, we will not know the true value of the population variance and we need to estimate it.

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

is a natural candidate for an estimate ...

But, ...

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

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

$$= \frac{1}{n}\sum_{i=1}^{n}\left(Var(X_i) + E(X_i)^2\right) - E(\bar{X}^2)$$

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

$$= \sigma^2 + \mu^2 - Var(\bar{X}) - (E(\bar{X}))^2$$

$$= \sigma^2 - \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \sigma^2\left(\frac{n-1}{n}\right)\frac{N}{N-1}$$

so $\hat{\sigma}^2$ is a biased estimate of σ^2.

In fact $\left(\frac{n-1}{n}\right)\frac{N}{N-1} < 1$

if $n < N$

so $E(\hat{\sigma}^2) < \sigma^2$.

An unbiased estimate of σ^2

is $\frac{N-1}{N} \cdot \frac{n}{n-1} \cdot \hat{\sigma}^2$.

$$= \frac{1}{n-1}\left(1 - \frac{1}{N}\right)\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Now since

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

we obtain an unbiased estimate of $Var(\bar{X})$ by replacing σ^2 with $\frac{N-1}{N}\,\frac{n}{n-1}\,\hat{\sigma}^2$

in the previous expression.

We get: the unbiased estimate

$$s_{\bar{X}}^2 \equiv \frac{\hat{\sigma}^2}{n}\left(\frac{n}{n-1}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-n}{N-1}\right)$$

$$= \frac{s^2}{n}\left(1-\frac{n}{N}\right)$$

where

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

It follows that

$$s_T^2 = N^2 s_{\bar{X}}^2$$

is an unbiased estimate of $Var(T)$.

In the special case where $X_i = 0$ or 1 for each i we have.

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

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

$$= \left(\frac{n}{n-1}\right) \left[\frac{1}{n} \left(\sum_{i=1}^{n} X_i^2\right) - \bar{X}^2 \right]$$

$$= \left(\frac{n}{n-1}\right) \left[\hat{p} - \hat{p}^2 \right].$$

$$= \frac{n}{n-1} \hat{p}(1-\hat{p})$$

By the above, we have

$$s_{\hat{p}}^2 = \frac{1}{n-1} \hat{p}(1-\hat{p}) \left(1 - \frac{n}{N}\right)$$

as an unbiased estimate of $Var(\hat{p})$.

Normal Approximation to the distribution of $\bar{X}$

If $X_1, \dots, X_n$ are i.i.d. with mean μ and variance σ^2

$\boxed{\text{Let } \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i}$

then

$$E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu.$$

$$Var(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}.$$

and C.L.T $\Rightarrow$

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le x \right) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}\, du.$$

as $n \to \infty$. $\left(\Phi(x) \right)$

In simple random sampling, the X_i are not independent but $Cov(X_i, X_j) = \frac{-\sigma^2}{N-1}$ $i \neq j$

is small for large N

and when ("$<<$" means "much less")

$$1 << n < N.$$

an approximate C.L.T applies.

e.g. when $1 << n << N$

we may approximate

$$P(|\bar{X}-\mu| \le \delta)$$

$$= P(-\delta \le \bar{X}-\mu \le \delta)$$

$$= P\left(-\frac{\delta}{\sigma_{\bar{X}}} \le \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \le \frac{\delta}{\sigma_{\bar{X}}}\right).$$

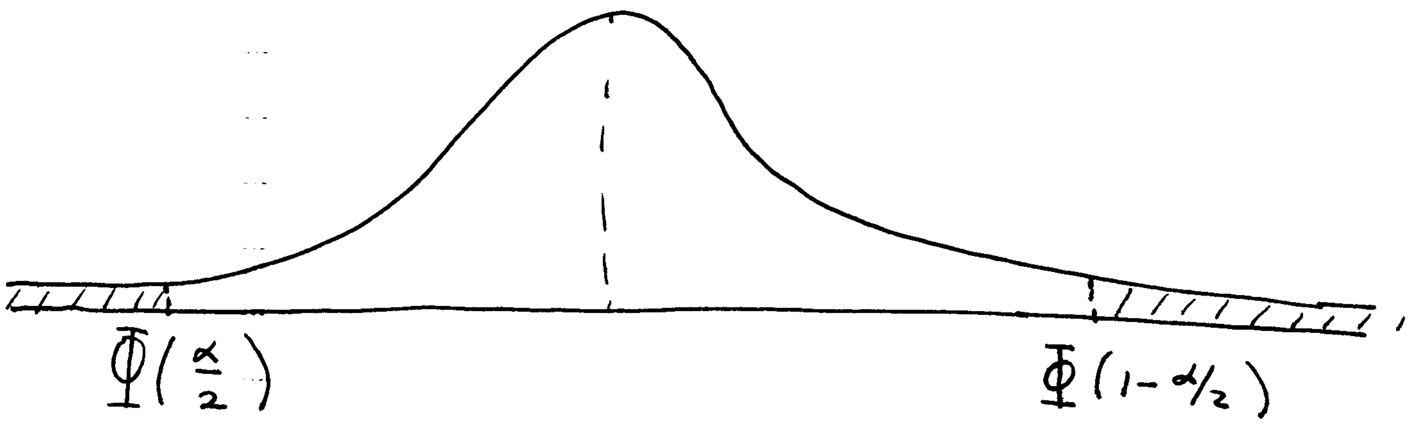
recall that

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$

$$= \Phi\left(\frac{\delta}{\sigma_{\bar{X}}}\right) - \Phi\left(-\frac{\delta}{\sigma_{\bar{X}}}\right) = 2\Phi\left(\frac{\delta}{\sigma_{\bar{X}}}\right) - 1.$$

In practice, we will not know σ^2 (or N) and we will estimate $\sigma_{\bar{X}}^2$ by
$$S_{\bar{X}}^2 \left(\text{or } \frac{s^2}{n}\right)$$

Confidence Intervals



If X is $N(0,1)$ then

$$\text{Prob}\left\{ \Phi\left(\frac{\alpha}{2}\right) \le X \le \Phi\left(1-\frac{\alpha}{2}\right)\right\} = 1-\alpha.$$

Note that $\Phi\left(\frac{\alpha}{2}\right) = -\Phi\left(1-\frac{\alpha}{2}\right)$.

Using our normal approximation to the distribution of $\bar{X}$, we assert that

$$P\left(\Phi\left(\tfrac{\alpha}{2}\right) \le \frac{\bar{X}-\mu}{s_{\bar{X}}} \le \Phi\left(1-\tfrac{\alpha}{2}\right) \right)$$

$(-\Phi(1-\tfrac{\alpha}{2}))$.

$$\approx 1-\alpha.$$

i.e.

$$P\left(\bar{X} - s_{\bar{X}}\Phi(1-\tfrac{\alpha}{2}) \le \mu \le \bar{X} + s_{\bar{X}}\Phi(1-\tfrac{\alpha}{2}) \right).$$

$$\approx 1-\alpha.$$

We call

$$\left(\bar{X} - \Phi(1-\tfrac{\alpha}{2})s_{\bar{X}} ,\ \bar{X} + \Phi(1-\tfrac{\alpha}{2})s_{\bar{X}} \right)$$

a $(1-\alpha)\%$ confidence interval for μ.

Notice that it is the interval which is random and that μ is a fixed value.

Computer examples.

- confidence intervals
- example D pg 219.

Estimation of a Ratio

If X is a random variable, and we know something about its 1st two moments $E(X)$ and $E(X^2)$ then we can sometimes say something useful about the 1st two moments of

$$Y = g(X)$$

if the function g is not badly behaved.

We expand g in a Taylor series about μ_X

$$Y = g(X) \simeq g(\mu_X) + (X-\mu_X)g'(\mu_X) + \frac{(X-\mu_X)^2}{2} g''(\mu_X) + \ldots$$

and from this we get.

$$E(Y) \simeq g(\mu_X) + \frac{g''(\mu_X)}{2} Var(X).$$

$$Var(Y) \simeq g'(\mu_X)^2 \, Var(X).$$

We know (by Tchebychev's inequality) that X will tend to be near μ_X.

~~If it is close enough +~~

If X is usually (i.e. w/ large prob.)

in a nbhd of μ_x for which g is well approximated by the 1st two terms of its Taylor expansion, then these approximations will be reasonably good.

Similarly, if X, Y are r.v's about whose 1st two moments, we have some information then we may argue as follows. for.

$$Z = g(X, Y).$$

$$Z = g(X, Y)$$

$$\approx g(\mu_X, \mu_Y) + (X-\mu_X)\frac{\partial g}{\partial x}(\mu_x, \mu_y)$$

$$+ (Y-\mu_Y)\frac{\partial g}{\partial y}(\mu_x, \mu_y)$$

$$+ \frac{1}{2}(X-\mu_x)^2 \frac{\partial^2 g}{\partial x^2}(\mu_x, \mu_y)$$

$$+ (X-\mu_X)(Y-\mu_Y)\frac{\partial^2 g}{\partial x \partial y}(\mu_x, \mu_y)$$

$$\vec{\mu} = (\mu_x, \mu_y) \qquad + \frac{1}{2}(Y-\mu_Y)^2 \frac{\partial^2 g}{\partial y^2}(\mu_x, \mu_y)$$

$$+ \ldots$$

to obtain

$$E(Z) \approx g(\mu) + \frac{1}{2}\sigma_X^2 \frac{\partial^2 g}{\partial x^2}(\vec{\mu}) + \frac{1}{2}\sigma_Y^2 \frac{\partial^2 g}{\partial y^2}(\vec{\mu})$$

$$+ \sigma_{XY}\frac{\partial^2 g(\vec{\mu})}{\partial x \partial y}$$

$$\mathrm{Var}(Z) \approx \sigma_X^2\left(\frac{\partial g}{\partial x}(\vec{\mu})\right)^2 + \sigma_Y^2\left(\frac{\partial g}{\partial y}(\vec{\mu})\right)^2 + 2\sigma_{XY}\left(\frac{\partial g}{\partial x}(\vec{\mu})\right)\left(\frac{\partial g}{\partial y}\vec{\mu}\right)$$

The application we have in mind is to.

$$Z = \frac{Y}{X} = g(X, Y).$$

for which we have.

$$\frac{\partial g}{\partial x} = -\frac{y}{x^2} \qquad \frac{\partial g}{\partial y} = \frac{1}{x}$$

$$\frac{\partial^2 g}{\partial x^2} = \frac{2y}{x^3} \qquad \frac{\partial^2 g}{\partial y^2} = 0.$$

$$\frac{\partial^2 g}{\partial x \partial y} = -\frac{1}{x^2}.$$

If $\mu_X \neq 0$ then we get.

$$E(Z) \simeq \frac{\mu_Y}{\mu_X} + \sigma_X^2 \frac{\mu_Y}{\mu_X^3} - \frac{\sigma_{XY}}{\mu_X^2}$$

$$= \frac{\mu_Y}{\mu_X} + \frac{1}{\mu_X^2}\left(\sigma_X^2 \frac{\mu_Y}{\mu_X} - \rho \sigma_X \sigma_Y\right)$$

$\rho = \text{corr}(X, Y)$.

also

$$\mathrm{Var}(Z) \simeq \sigma_X^2 \frac{\mu_Y^2}{\mu_X^4} + \frac{\sigma_Y^2}{\mu_X^2} - 2\sigma_{XY}\frac{\mu_Y}{\mu_X^3}$$

$$= \frac{1}{\mu_X^2}\left(\sigma_X^2 \frac{\mu_Y^2}{\mu_X^2} + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y\frac{\mu_Y}{\mu_X}\right)$$

Note that

$\left|E(Z) - \frac{\mu_Y}{\mu_X}\right|$ can be large if

μ_X is small or if

σ_X, σ_Y are large.

also.

$\mathrm{Var}(Z)$ is large if μ_X is small

but correlation between X, Y which

is of the same sign as $\frac{\mu_Y}{\mu_X}$

decreases the variance.