1 Lecture 6 Monday 01/29/01

Homework and Sectionning: see the logictics section

The hw set does not include problem 50, but does include problem 51.

There are NO computer labs this week. There will be computer next week (week of Feb.5th).

I put the 2 I had lost in red

7.4 Review of last time, if you missed it, it was *hard*, see: Lecture 5

7.5 Estimating a Ratio

Suppose now that we have a population of N pairs $(x_1, y_1), (x_2, y_2) \dots (x_N, y_N)$ The ratio of interest is

$$r = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x}$$

$$R = \frac{\bar{Y}}{\bar{X}}$$

7.6 Approximation Methods

Chebychev's Theorem :

$$P(|X - \mu_X| > k\sigma_X) < \frac{1}{k^2}$$

$$Z = g(X, Y) = g(\mu) + (X - \mu_X)\frac{\partial g}{\partial x}(\mu) + (Y - \mu_Y)\frac{\partial g}{\partial y}(\mu)$$

$$Var(Z) = \frac{\partial g}{\partial x}(\mu)^2\sigma_X^2 + \frac{\partial g}{\partial y}(\mu)^2\sigma_Y^2 + 2\frac{\partial g}{\partial x}(\mu)\frac{\partial g}{\partial y}(\mu)\sigma_{X,Y}$$

$$\begin{aligned} Z = g(X, Y) = \quad & g(\mu) + (X - \mu_X)\tfrac{\partial g}{\partial x}(\mu) + (Y - \mu_Y)\tfrac{\partial g}{\partial y}(\mu) \\ & + \tfrac{1}{2}(X - \mu_X)^2\tfrac{\partial^2 g}{\partial x^2}(\mu) + \tfrac{1}{2}(Y - \mu_Y)^2\tfrac{\partial^2 g}{\partial y^2}(\mu) \\ & + (X - \mu_X)(Y - \mu_Y)\tfrac{\partial^2 g}{\partial x \partial y}(\mu) \end{aligned}$$

From this and from the properties of the expectation:

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

7.6.1 Expectation of the mean and variance of a ratio

$$g(x,y) = \frac{y}{x} \text{and } Z = \frac{Y}{X}$$

$$\frac{\partial g}{\partial x} = \frac{-y}{x^2} \quad \frac{\partial g}{\partial y} = \frac{1}{x}$$
$$\frac{\partial^2 g}{\partial x^2} = \frac{2y}{x^3} \quad \frac{\partial^2 g}{\partial y^2} = 0$$
$$\frac{\partial^2 g}{\partial x \partial y} = -\frac{1}{x^2}$$

Supposing that $\mu_x \neq 0$

$$E(Z) = \frac{\mu_y}{\mu_x} + \frac{1}{2}\sigma_X^2 2\frac{\mu_y}{\mu_x^3} - \sigma_{XY}\frac{1}{\mu_x^2}$$

and

$$Var(Z) = \frac{\mu_y^2}{\mu_x^4}\sigma_X^2 + \frac{1}{\mu_x^2}\sigma_Y^2 - \frac{\mu_y}{\mu_x^3}\sigma_{XY}$$

Now, we need to adapt these formulas to the case we need here which is where the random variables X and Y are $\bar{X}$ and $\bar{Y}$.

We will also use a population parameter we did not define in the first lecture but it is easy to guess it's definiton by it's name:

Population covariance

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

and while I'm adding population parameters, let's define the population correlation coefficient:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

We use it because it can be proved by the same argument I used in Th 3.1B that :

$$cov(\bar{X}, \bar{Y}) = \frac{\sigma_{xy}}{n}(1 - \frac{n-1}{N-1})$$

Theorem 1 (7.4.A)

$$Var(R) \approx \frac{1}{\mu_x^2}(r^2\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 - 2r\sigma_{\bar{X}\bar{Y}}) = \frac{1}{\mu_x^2}\frac{1}{n}(1 - \frac{n-1}{N-1})(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y)$$

A strong correlation between x and y will decrease the variance of R.

Beware: we still have a problem when $\bar{X}$ is small, the variance will then be large, small values in $\bar{X}$ make R vary badly.

The covariance is estimated by

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

so an estimate of the variance of R is thus:

$$\begin{aligned} \hat{Var}(R) &= \frac{1}{\mu_x^2}(r^2\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 - 2r\sigma_{\bar{X}\bar{Y}}) \\ &= \frac{1}{\mu_x^2}\frac{1}{n}(1 - \frac{n-1}{N-1})(r\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y) \end{aligned}$$

From above we translate for the expectation:

Theorem 2 (7.4.B) *With simple random sampling, the expectation of R is approximated by :*

$$E(R) \doteq r + \frac{1}{\mu_x^2}\frac{1}{n}(1 - \frac{n-1}{N-1})(r\sigma_x^2 - \rho\sigma_x\sigma_y)$$

At this point I would like to say a word more about variance and bias, you may or may not have heard of the mean squared error, what is it for an estimate θ_2 ?

$$MSE = E[(\hat{\theta} - \theta)^2]$$

Of course this is something we want to minimize and if we knew an optimum, a minimizer we should use it, but we don't usually have one.

This average squared deviation can be broken down into two pieces:

$$MSE = E[(\hat{\theta} - E(\hat{\theta}))^2] + [E(\hat{\theta}))^2 - \theta]^2$$

or said otherwise :

$$MSE = var(\hat{\theta}) + [bias(\hat{\theta})]^2$$

In our case here, the variance is of order $\frac{1}{n}$, but so is the bias, so the bias's contribution to the MSE could be neglected if the sample size is large enough.

Minimizing the variance is what matters here.

This is case when we are going to sacrifice bias in the variance/bias tradeoff because all in all that will improve the MSE.

Now of course all this was theoretical because again it gives orders of magnitude but if we want real numbers we will have to replace everywhere the unknown parameters by their estimates from OUR given unique sample.

As the random variables of which we are taking ratios have Normal distributions because of teh central limit theorem, and because of our linearization to first order R is a linear combinations of Normals, it will thus be Normal.

This is useful, as before , for constructing confidence intervals for r using the Normal, we need of course an estimate of the standard error, this is provided by theorem 4.A. Useless

as it stands, everything therein has to be estimated, we plug in all the estimates we have to get:

$$S_R^2 = \frac{1}{\bar{X}^2}\frac{1}{n}(1 - \frac{n-1}{N-1})(R^2 S_X^2 + S_Y^2 - 2RS_{XY})$$

and a $100(1-\alpha)\%$ confidence interval is given by :$R \pm z_{\frac{\alpha}{2}} S_R$ Example A (209)

7.6.2 A Ratio as a tool for estimating a mean

In the hospital beds problem, it could have been that the number of discharges was unknown, but the precise size of the hospitals in numbers of beds was known from an earlier enumeration. Call this known quantity μ_x, then we can estimate μ_y by

$$\bar{Y}_R = \mu_x \frac{\bar{Y}}{\bar{X}} = \mu_x R$$

We expect X and Y to be heavily correlated, this can be actually checked on the complete population.

If we have a sample where $\bar{X} < \mu_x$ the sample underestimates the # beds and probably the number of discharges as well.

Multiplying $\bar{Y}$ by the proportionality factor $\frac{\mu_x}{\bar{X}}$ increases $\bar{Y}$ to $\bar{Y}_R$.

Now we will look first to empirical evidence about the idea :the simulation was of B=500 samples of size n=64 from the population of N=393.

Here is an algorithm that would do a simulation of B=500 samples of size 64 from a population vector called pop of length 393:

```
function out=srs(Npop,n){
%Outputs a vector of indices for simple
%random sampling
tmp=randperm(Npop);
out=tmp(1:n);
}
function out=meanr(datax,datay,mux)
{%Ratio estimate of the mean
out=mux*(mean(datay)/mean(datax))
}
%Ordinary means
results=zeros(1,500)   % initializing step
for ( s = (1:500)) {   % S loop
    sampleb=srs(393,64)
    results(b)=mean(datapopy(sampleb))
    }                  % End of  S loop
%Ratio means
results=zeros(1,500)   % initializing step
mux=mean(datapopx);
for ( s = (1:500)) {   % S loop
    sampleb=srs(393,64);
    results(b)=meanr(datapopx(sampleb),datapopy(sampleb),mux);
    }                  % End of  S loop
%Ratio means
```

Now if we needed to do this several times we would be better off creating a function: