Math 152 ~~1/29/09~~ 2/3/09

Estimation of a Ratio. (Previous notes also contained 1/29/09)

Suppose that for each population member we have two values (x_i, y_i)

e.g. a) i^{th} hospital.

(y_i = discharges, x_i = beds).

b) i^{th} county.

(y_i = cancer mortality, x_i = white female population).

We will try to "measure" or estimate

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

$$\left(\neq \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i}{x_i} \right). \right)$$

e.g. a) r = #patients discharged per bed.

b) r = total proportion of cancer mortality among white females.

Drawing a sample of pairs (X_i, Y_i)

the natural estimate for r is

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

and we would like information on

$E(R)$, $Var(R)$ and (generally) the distribution of R.

We can get approximate information using the results of the previous discussion.

We will need

$Var(\bar{X})$, $Var(\bar{Y})$ (which we know)

and $Cov(\bar{X}, \bar{Y})$.

Let

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

(population covariance).

We claim that

(*) $$\mathrm{Cov}(\bar{X}, \bar{Y}) = \frac{\sigma_{xy}}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Using our previous formulas, this gives us. (for $R = \frac{\bar{Y}}{\bar{X}}$)

$$\mathrm{Var}(R) \approx \frac{1}{\mu_x^2}\left(r^2\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 - 2r\sigma_{\bar{X}\bar{Y}}\right)$$

$$= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}\right)$$

and

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

$$\left(\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}\right).$$

Before considering these approximate expressions, let's see why (*) holds.

note that

$$\langle x-y, x-y\rangle = \langle x, x\rangle - 2\langle x, y\rangle + \langle y, y\rangle$$

so $$\langle x, y\rangle = \frac{\langle x, x\rangle + \langle y, y\rangle - \langle x-y, x-y\rangle}{2}.$$

$$\therefore \quad \mathrm{Cov}(\bar{X}, \bar{Y}) = \frac{\mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) - \mathrm{Var}(\bar{X}-\bar{Y})}{2}.$$

$$= \frac{1}{2}\left(\frac{\sigma_x^2}{n}\left(1 - \frac{n-1}{N-1}\right) + \frac{\sigma_y^2}{n}\left(1 - \frac{n-1}{N-1}\right) - \frac{\sigma_{(x-y)}^2}{n}\left(1 - \frac{n-1}{N-1}\right)\right).$$

Now $$\sigma_x^2 + \sigma_y^2 - \sigma_{(x-y)}^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[(x_i-\mu_x)^2 + (y_i-\mu_y)^2 - (x_i - y_i - (\mu_x-\mu_y))^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[(x_i-\mu_x)^2 + (y_i-\mu_y)^2 - \left((x_i-\mu_x) - (y_i-\mu_y)\right)^2\right]$$

$$= 2\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y).$$

So $$\mathrm{Cov}(\bar{X}, \bar{Y}) = \frac{\sigma_{xy}}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Now

$$\mathrm{Var}(R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y\right),$$

$$\left(\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}\right).$$

and $$E(R) \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r\sigma_x^2 - \rho\sigma_x\sigma_y\right).$$

- If ρ is the same sign as r, this tends to lower the variance, and the bias.

- The contribution of the bias to the mean squared error (in approximating r by R) is on the order of $\frac{1}{n^2}$, while the contribution of the variance is on the order of $\frac{1}{n}$. (recall: $\mathrm{mse} = \mathrm{variance} + (\mathrm{bias})^2$)

Since R is (for large samples) approximated by a linear combination of $\bar{X}-\mu_X$ and $\bar{Y}-\mu_Y$, R will be approximately normally distributed and we can find approximate confidence intervals for r.

We have (~~from last time~~).

$$\text{Var}(R) \approx \frac{1}{\mu_{\bar{X}}^2}\left(r^2\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 - 2r\sigma_{\bar{X}\bar{Y}}\right)$$

$$= \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\frac{1}{\mu_X^2}\left(r^2\sigma_X^2+\sigma_Y^2-2r\sigma_{XY}\right)$$

Since we don't know μ_X, r, σ_X, σ_Y or σ_{XY}. we use instead (respectively)

$\bar{X}$, R, s_X^2, s_Y^2 and $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$.

so the estimated variance of R is

$$s_R^2 = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\bar{x}^2}\left(R^2 s_x^2 + s_y^2 - 2R s_{xy}\right)$$

An interval which contains r with approximate probability $(1-\alpha)$ is

$$R \pm \Phi^{-1}(1-\alpha/2)\, s_R.$$

We now apply the previous ideas to reduce variance in the estimation of a sample mean.

e.g. x_i = # of beds in i^{th} hospital.

y_i = # of discharges in i^{th} hospital.

We suppose that μ_x is known from an earlier survey and consider the <u>ratio estimate</u> of μ_y.

$$\bar{y}_R = \frac{\mu_x}{\bar{x}}\bar{y} = \mu_x R.$$

Why should this estimate perform better?
(than $\bar{Y}$).

We know that there will be some bias
but we expect a reduction in variance
because of the expected correlation in x_i and y_i.
Intuitively If $\bar{X} < \mu_x$ then the
sample underestimates the # of beds and
~~we expect # of beds to be correlated with~~
~~# of discharges. so~~ probably also
underestimates the number of discharges
Multiplying $\bar{Y}$ by $\frac{\mu_x}{\bar{X}}$ pushes the
estimate upward (in probably the right direction).
Similarly if $\bar{X} > \mu_x$ then
multiply $\bar{Y}$ by $\frac{\mu_x}{\bar{X}}$ brings
down the estimated value $\bar{Y}$.

From our previous formulas we have.

$$\mathrm{Var}(\bar{Y}_R) \approx \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\left(r^2\sigma_x^2+\sigma_y^2-2r\rho\sigma_x\sigma_y\right)$$

and $$E(\bar{Y}_R)-\mu_y \approx \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\frac{1}{\mu_x}\left(r\sigma_x^2-\rho\sigma_x\sigma_y\right)$$

~~Bias~~ (Note that the ~~standard error~~ bias is on the order of the square of the standard error).

Since $\mathrm{Var}(\bar{Y}) \approx \frac{\sigma_y^2}{n}$.

we can expect a reduction in variance when

$$r^2\sigma_x^2 - 2r\rho\sigma_x\sigma_y < 0.$$

Supposing $r>0$, this is

$$2\rho\sigma_y > r\sigma_x = \frac{\mu_y}{\mu_x}\sigma_x$$

So we need

$$\rho > \frac{1}{2}\left(\frac{\sigma_x/\mu_x}{\sigma_y/\mu_y}\right).$$

$$C_x = \frac{\sigma_x}{\mu_x} \qquad C_y = \frac{\sigma_y}{\mu_y}.$$

are called coefficients of variation

$\bar{Y}_n$ is also approximately normally distributed and (from our earlier formula).

$$S_{\bar{Y}_n}^2 = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\left(R^2 S_x^2 + S_y^2 - 2R S_{xy}\right)$$

So an ~~approximate~~ interval containing μ_y with approximate prob. $1-\alpha$ is

$$\left(\bar{Y}_n \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right) S_{\bar{Y}_n}\right).$$

e.g. From our hospitals example we have

$$\mu_x = 274.8 \qquad \sigma_x = 213.2$$
$$\mu_y = 814.6 \qquad \sigma_y = 589.7$$
$$r = 2.96 \qquad \rho = .91.$$

$$\text{Var}(\bar{Y}_R) \approx \frac{1}{n}\left(r^2\sigma_x^2 + \sigma_y^2 - 2 \cdot r \cdot \rho \cdot \sigma_x \cdot \sigma_y\right)$$

$$\approx \frac{68,697.4}{n}.$$

$$\sigma_{\bar{Y}_R} \approx \frac{262.1}{\sqrt{n}}.$$

With $n = 64$ and including the finite population correction we get

$$\sigma_{\bar{Y}_R} \approx \frac{262.1}{8}\sqrt{1 - \frac{63}{392}} \approx 30.0$$

Note that the standard deviation of $\bar{Y}$ is

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}} \approx \frac{589.7}{8}\sqrt{1 - \frac{63}{329}}$$

$$\approx 66.3.$$

Notice that a ratio estimate with a sample of size n_1 has.

$$\bar{\theta}_{R} \approx \frac{262.1}{\sqrt{n_1}}$$

while an simple estimate with sample size n_2 has.

$$\theta_{\bar{Y}} \approx \frac{589.7}{\sqrt{n_2}}$$

and to get the variances to be the same we need.

$$n_1 = n_2 \left(\frac{262.1}{589.7} \right)^2 \approx .1975\, n_2.$$

So we get the same precision from the ratio estimate with $\approx 1/5$ the sample size.