- an argument based on the CLT can be used to show that R is approximately normally distributed, i.e., $R \overset{D}{\approx} N\left(\mu_R, \sigma_R^2\right)$, when sample size n is large.
- **Applications** ($\because$ bias$\to$0) r_{xy} ← $n\to\infty$; σ_R^2 can be replaced by S_R^2 ($\because S_R^2 \xrightarrow{P} \sigma_R^2$)
  - probability of estimation error $\in [a,b]$, e.g., $P\left(\left|\frac{R-r_{xy}}{s_R}\right| > \delta\right) \approx 2[1-\Phi(\delta)]$ ($N(0,1) \overset{D}{\approx}$ …)
  - approximate $100(1-\alpha)\%$ confidence interval of r_{xy}: $R \pm z(\alpha/2)\, s_R$

10/2 ↓

Example 16 (estimate population ratio r_{xy})

- Suppose that 100 people who recently bought houses are surveyed, and
  y: mortgage payment x: gross income
  are observed. The $r_{xy} = \tau_y/\tau_x$ is the percentage of the total mortgage amount to the total gross income of all people who recently bought houses.
- Suppose that the population size N is missing, but it is known that $100 \ll N$. (without ≈ with)
- Suppose that $\overline{X} = 3100$, $s_x = 1200$ ($\overset{e}{\to}\sigma_x$), $\overline{Y} = 868$, $s_y = 250$, $\hat{\rho}_{xy} = 0.85$. (treated as s.r.s. with repl.)
  We have $R = 868/3100 = 0.28$. ($\overset{e}{\to}\mu_x$, $\overset{e}{\to}\mu_y$, $\overset{e}{\to}\sigma_y$, $\overset{e}{\to}\rho_{xy}$; cf. r_{xy}; LNp.42)
- Neglecting the finite population correction, the estimated standard error of R is (Def 14. S.r.s. with (LNp.50); shows accuracy of R)
  $$s_R = \frac{1}{10} \times \frac{1}{3100} \sqrt{0.28^2(1200^2) + 250^2 - 2(0.28)(0.85)(250)(1200)} = 0.006.$$
  Note that s_R is small because ① x and y are highly positively correlated, $r_{xy} > 0$, and ② $\overline{X}$ is large. ← check the graph in LNp.49

- An approximate 95% confidence interval for r_{xy} is
  $$0.28 \pm 1.96 \times 0.006 = 0.28 \pm 0.012 = (0.268, 0.292).$$ ← an interval estimate
- Again, neglecting the finite population correction, an estimated bias of R using Thm 16 (LNp.48) is (① $\mathrm{Var} \sim O(n^{-1})$ cf. $O(n^{-2}) \sim \mathrm{bias}^2$)
  $$\frac{1}{n} \times \frac{1}{\overline{X}^2}\left(Rs_x^2 - \hat{\rho}_{xy} s_x s_y\right) = \frac{1}{100} \times \frac{1}{3100^2}\left[(0.28)(250^2) - (0.85)(250)(1200)\right] = -0.00025,$$
  (same sign; ②)
  which is negligible relative to s_R (=0.006). Note that the large $\hat{\rho}_{xy}$ (=0.85) and the large value of $\overline{X}$ (=3100) cause the bias to be small. (③)

(s_R in LNp.51 ↔ cf.)

• Ratios used for estimating population means (and totals)

- Suppose μ_x is known, e.g., the example of 393 hospitals in Ex.2 (LNp.4),
  - y: number of discharges, ← main interest μ_y
  - x: number of beds.
  (population (N is known); same type of data as in estimation of a ratio)

  Suppose the average (or total) number of beds μ_x (or τ_x) in the 393 hospitals is known (before a sample has been taken). (Why multiply $\overline{Y}$ by this? Check the explanation in LNp.54)
- Q: how to take advantage of this information in the estimation of μ_y?
- Select a random sample, and collect the data: (X_k, Y_k), $k = 1, \ldots, n$. For the parameter $\mu_y = \mu_x r_{xy}$, an intuitive **ratio estimator** of μ_y is (Both X_k and Y_k are r.v.'s; use (x,Y) data to estimate μ_y)
  $$\overline{Y}_R = \mu_x R = \overline{Y}\left(\frac{\mu_x}{\overline{X}}\right)$$ (only use Y data to estimate μ_y)
  ($\overset{\text{cf.}}{\longleftrightarrow} \overline{Y}$; Q: which estimator of μ_y is better?).

- In the following discussion of this topic, we only consider the case of s.r.s. without replacement. The case of s.r.s. with replacement follows analogously. → remove finite population correction

Example 17 (Comparison of sample mean and ratio estimator, cont. Ex.2 in LNp.4)

- Consider the example of hospital discharges. For the population of 393 (N) hospitals and $1 \le i \le N$, let
  - x_i: number of beds in the ith hospital (known before sampling)
  - y_i: number of discharges in the ith hospital

  positively highly correlated

  Observe Data $(X_k, Y_k), k=1,\cdots,n$ random, too.

- In this population,

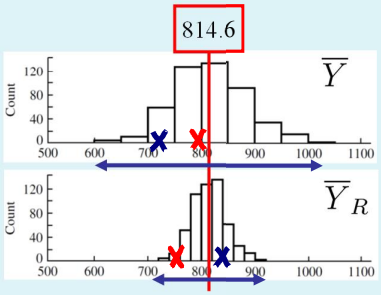
  $\mu_x = 274.8$ (known),
  $\sigma_x = 213.2$ (known),
  $\mu_y = 814.6$, $\sigma_y = 589.7$,
  $r_{xy} = 2.96$, $\rho_{xy} = 0.91$.

  ×: unknown in sampling survey

  parameters

  For (X_k, Y_k)'s, same statistical modeling as given in LNp.41

  Note. not $|\bar{Y} - \mu_y| > |\bar{Y}_R - \mu_y|$



- To compare the preformance of $\bar{Y}$ and $\bar{Y}_R$, it was simulated (check Ex.3, LNp.15) 500 samples of size 64 (n) from the population of hospitals.

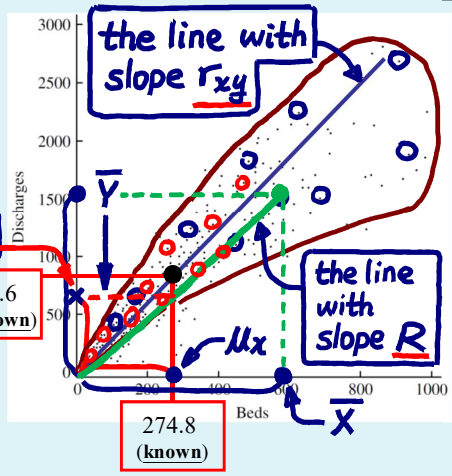
  Sampling distributions

- The histograms of this result are shown in Figure 7.6 of textbook.
  - The histograms show that the ratio estimator $\bar{Y}_R$ of μ_y is less variable than the sample mean $\bar{Y}$.

  MSE($\bar{Y}$) > MSE($\bar{Y}_R$)

  - The comparison shows the ratio estimator $\bar{Y}_R$ is effective at reducing variability $\Rightarrow$ $\bar{Y}_R$ is a more accurate estimator than $\bar{Y}$. ← always true ?

- Q: Why is $\bar{Y}_R$ better than $\bar{Y}$ in this case?
- An explanation. Check the scatterplot of (x_i, y_i) for the 393 hospitals in the population (Figure 7.5 of textbook) and consider a random sample (X_k, Y_k), $k = 1, \ldots, n$.



  - the population correlation $\rho_{xy} = 0.91$ is high $\Rightarrow$ a hospital with a large x_i tends to have a large y_i

  What if $\bar{X} < \mu_x$?

  if $\bar{X} > \mu_x$, the sample over-estimates the number of beds μ_x, and probably the number of discharges as well, i.e., probably $\bar{Y} > \mu_y$. ← ∵ extra information

  r.v.

  錨點 μ_x ↔ $\bar{X}$ cf ↕ cf $\bar{Y}$ ⇐ μ_y = ?

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

  - for this sample, multiplying $\bar{Y}$ by $\frac{\mu_x}{\bar{X}}$ decreases $\bar{Y}$ to $\bar{Y}_R$, which might be closer to μ_y than $\bar{Y}$.

  $\frac{\mu_x}{\bar{X}} < 1$; $\bar{Y}(\mu_x/\bar{X})$

Theorem 19 (approximate mean, bias, and variance of the ratio estimator)

Since $\bar{Y}_R = \mu_x R$, we have $E(\bar{Y}_R) = \mu_x E(R)$ and $Var(\bar{Y}_R) = \mu_x^2 Var(R)$. Under s.r.s. without replacement,

a known constant

Thm16 (LNp.48)

not unbiased

Thm17 (LNp.49)

- the approximate bias of the ratio estimator $\bar{Y}_R$ of μ_y is

$$E(\bar{Y}_R) - \mu_y = \mu_x [E(R) - r_{xy}] \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right) \times \frac{1}{\mu_x}\left(r_{xy}\sigma_x^2 - \rho_{xy}\sigma_x\sigma_y\right),$$

Thm16 (LNp.48)

Bias(R) ; μ_y/μ_x

Bias($\bar{Y}_R$)^2 ~ $O(n^{-2})$

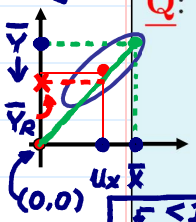
- The approximate variance of the ratio estimator $\overline{Y}_R$ of μ_y is

$$\sigma^2_{\overline{Y}_R} = Var(\overline{Y}_R) = \mu_x^2 Var(R) \approx \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\times(r_{xy}^2\sigma_x^2+\sigma_y^2-2r_{xy}\rho_{xy}\sigma_x\sigma_y).$$

Proof: The results follows directly from the formulas for the approximate mean and variance of R given in Thm. 16 (LNp.48) and Thm. 17 (LNp.49).

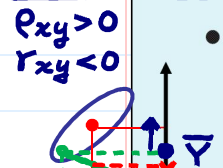
Note 13 (A note about the approximate variance of the ratio estimator)

Q: When will the ratio estimator $\overline{Y}_R$ be better than the ordinary estimator $\overline{Y}$, i.e., $Var(\overline{Y}_R) < Var(\overline{Y})$?



- The ordinary estimator $\overline{Y}$ has variance $\sigma^2_{\overline{Y}} = Var(\overline{Y}) = \frac{\sigma_y^2}{n}\left(1-\frac{n-1}{N-1}\right)$ (Thm. 3, LNp.18) and

$$\sigma^2_{\overline{Y}_R} - \sigma^2_{\overline{Y}} = Var(\overline{Y}_R) - Var(\overline{Y}) \approx \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\times(r_{xy}^2\sigma_x^2 - 2r_{xy}\rho_{xy}\sigma_x\sigma_y).$$



- The ratio estimator $\overline{Y}_R$ has a smaller variance than $\overline{Y}$ if

$$r_{xy}^2\sigma_x^2 - 2r_{xy}\rho_{xy}\sigma_x\sigma_y < 0 \quad \Leftrightarrow \quad r_{xy}^2\sigma_x < 2\, r_{xy}\,\rho_{xy}\,\sigma_y$$

i.e.

$$\rho_{xy} \begin{cases} > \frac{1}{2}\left(\frac{\mu_y}{\mu_x}\right)\left(\frac{\sigma_x}{\sigma_y}\right) = \frac{1}{2}\left(\frac{CV_x}{CV_y}\right) > 0, & \text{provided that } r_{xy} > 0, \\ < \frac{1}{2}\left(\frac{\mu_y}{\mu_x}\right)\left(\frac{\sigma_x}{\sigma_y}\right) = \frac{1}{2}\left(\frac{CV_x}{CV_y}\right) < 0, & \text{provided that } r_{xy} < 0. \end{cases}$$

where $CV_x = \sigma_x/\mu_x$ and $CV_y = \sigma_y/\mu_y$ are the **coefficients of variation**.