

STT 315

Week of October 9, 2006

We take up chapter 7 beginning the week of October 16.

This week 10-9-06 expands on chapter 6, after which you will be equipped with yet another powerful statistical idea enabling you to offer estimates at reduced sampling cost. This will be your first brush with the concept of **correlation** which gets at the issue of statistical dependency. Dependency is a major player in statistics since it allows us to leverage side information. It is the reason pollsters ask so many questions seemingly unrelated to the issue at hand.

Also new this week, you will learn to properly estimate differences $D = \mu_1 - \mu_2$ or $D = p_1 - p_2$ between means or proportions. With such methods you will be able to offer a statistical basis for choosing between such things as different advertising methods or different packaging designs based on estimates of the sales impact.

Usual CI half-width. The formula below is the familiar one appropriate to sampling with-replacement when the z-CI method is employed for large n. If we take $z = 1.96$ we get the ME for \bar{y} as an estimator of $E Y$ (i.e. $\mu(y)$).

$$\pm z \frac{s_y}{\sqrt{n}}$$

Smaller CI half-width. What if you could instead use the same sample size n to obtain

$$\pm z \frac{s_y}{\sqrt{n}} \sqrt{1 - \hat{\rho}^2}$$

for some quantity ρ HAT, called the sample correlation, whose square cannot exceed one? That would be good since it would mean a narrower CI for the same sample.

Catch to the better CI. To enjoy the narrower CI half-width above we must undertake a modification of the usual with-replacement sampling plan. It will require that we KNOW $\mu(x)$. To illustrate the idea suppose we've the job of estimating our mean sales to our accounts this year (y). We could just sample large n accounts and speak with their representatives to estimate their likely purchase amount y from us this year.

At this point we'd have scores Y_1, \dots, Y_n and the half-width (two formulas up) would apply if we use \bar{y} . But we are ignoring useful information that is freely available! It

would be an easy matter to look up x = amount purchased from us last year, for each sample account..

If we do that, we really have samples consisting of pairs (x, y)

$$(X_1, Y_1) \dots\dots (X_n, Y_n)$$

and these samples are independent from 1 to n , although (x, y) may be dependent within pairs.

So what is rhoHAT? It is the sample correlation coefficient between x and y , measuring the degree of linear dependency between x and y . On page 449 of your book there are several (x, y) plots given together with their correlations. Read pp.448-450 except the little part at the end (on testing).

$$\hat{\rho} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sqrt{\overline{X^2} - (\bar{X})^2} \sqrt{\overline{Y^2} - (\bar{Y})^2}}$$

So do we just use $y\text{BAR} \pm z (s_y / \text{root}(n)) \text{root}(1 - \text{rhoHAT}^2)$? Not quite! You haven't used the x -information to modify the estimator! Here is the new estimator you must use in place of $y\text{BAR}$ (let's call it the regression-based estimator)

$$\bar{Y} + (\mu_x - \bar{x}) \hat{\rho} \frac{s_y}{s_x}$$

it is equivalent to the following, i.e. $n-1$ divisors cancel from numerator and denominator

$$\bar{Y} + (\mu_x - \bar{x}) \hat{\rho} \frac{\sqrt{\overline{y^2} - (\bar{Y})^2}}{\sqrt{\overline{x^2} - (\bar{x})^2}}$$

How to interpret the above estimator of $\mu(y)$? First, you cannot use it unless you know the population mean $\mu(x)$! In our example it is ok since we would know the total, and hence the average $\mu(x)$, of the purchases from ALL OF OUR ACCOUNTS LAST YEAR. Second, the formula above seems do be doing a sensible thing. For example, if $x\text{BAR}$ falls below the known mean $\mu(x)$ that suggests that $y\text{BAR}$ will be also fall below $\mu(y)$ and we should boost our estimate up from $y\text{BAR}$. But that is exactly what the formula is doing in this case since it is boosting $y\text{BAR}$ by a positive multiple of $(\mu(x) - x\text{BAR})$.

How to interpret the CI half-width we are then entitled to use? Here it is again:

$$\pm z \frac{s_y}{\sqrt{n}} \sqrt{1 - \hat{\rho}^2}$$

It seems likely that the population correlation ρ between purchases x last year and y this year will be near one. So the sample correlation ρHAT should be near one also. Therefore $\sqrt{1 - \rho\text{HAT}^2}$ will be near zero! This will earn us far greater precision over simply using $y\text{BAR}$. And all for the pennies cost of looking up what our sample accounts spent with us last year!

Exercises due in recitation Thursday 12th.

1. Suppose that 50 customers are sample with replacement and scored for x = amount they purchased with us last year, y = amount we determine they will likely purchase from us this year (after examining their situation and consulting with the account rep). We wish to estimate $\mu(y) = E Y$ = average y -score for the entire population of our thousands of accounts. Suppose the sample data for these 50 accounts gives

$$\begin{array}{ll} x\text{BAR} = 9843 & y\text{BAR} = 8810 \\ s(x) = 480 & s(y) = 527 \\ \rho\text{HAT} = 0.694 \end{array}$$

Suppose that we KNOW the averages sales amount for ALL of our accounts was 9900 last year (i.e. $\mu(x) = 9900$).

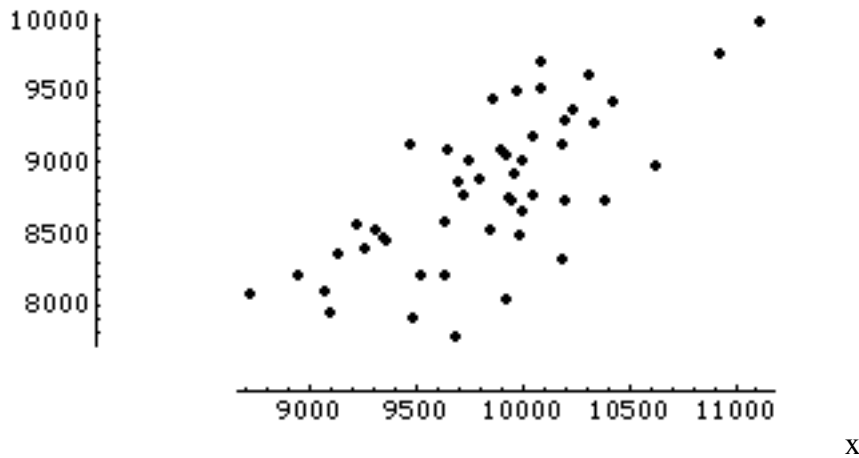
a. Give the usual 95% CI for $\mu(y)$ based on the with replacement sample of 50 y -scores.

b. Compare the estimate $y\text{BAR}$ with the regression-based estimate.

c. Give the 95% CI for $\mu(y)$ based on the regression approach.

d. If it cost \$400 to evaluate each sample y how much money would it cost to achieve our greater precision (c) by just increasing n ? (see what n in (a) would do it).

e. Here is the sample of 50.



The SAMPLE REGRESSION LINE passes through the point (\bar{x}, \bar{y}) with slope

$$\text{slope} = \hat{\rho} \frac{s_y}{s_x}$$

Calculate this slope and draw the sample regression line in the above plot.

f. Since the regression line passes through (\bar{x}, \bar{y}) it may be obtained visually by entering \bar{x} to the horizontal axis and reading off \bar{y} (the estimate of $\mu(y)$) from the line. Likewise obtain the regression based estimator of $\mu(y)$ when I tell you that it is obtained by inserting KNOWN $\mu(x)$ into the horizontal axis and reading off the height of the sample regression line at that point. DO IT AND SEE.

The idea behind the regression based estimator of $\mu(y)$ is that since the sample line is an estimate of the population line, and the population line passes through $(\mu(x), \mu(y))$, we should put $\mu(x)$ (known) into the sample line.

2. **CI for difference of means with paired data.** Read table 8-1 pg. 326 and CI (8-3) pg. 330 for the difference D between two means WITH PAIRED DATA. Using the data of 8-1 confirm the calculation below (8-3).

3. Solve problem 8-2 (paired data).

4. **CI for difference of means with independent samples.** Read the first two paragraphs of pg. 333 and the expression at the very bottom of the page. Read the Confidence Intervals part of pg. 338. The z-based CI for n_1 and n_2 large and the population standard deviations unknown is

$$\pm z \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

Solve 8-22.

5. Read the Confidence Intervals material beginning at the bottom of pg. 349. Solve problem 8-30.