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 <u>differences</u> 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 yBAR 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 rhoHAT, 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, ... Y_n and the half-width (two formulas up) would apply if we use yBAR. 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) ..... (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).

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

So do we just use yBAR +/- z (s_y / root(n)) root(1 – 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 yBAR (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

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

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 xBAR falls below the known mean mu(x) that suggests that yBAR will be also fall below mu(y) and we should boost our estimate up from yBAR. But that is exactly what the formula is doing in this case since it is boosting yBAR by a positive multiple of (mu(x) – xBAR).

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 rho between purchases x last year and y this year will be near one. So the sample correlation rhoHAT should be near one also. Therefore root(1 – rhoHAT^2) will be near zero! This will earn us far greater precision over simply using yBAR. 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

xBAR = 9843 yBAR = 8810
s(x) = 480 s(y) = 527
rhoHAT = 0.694

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

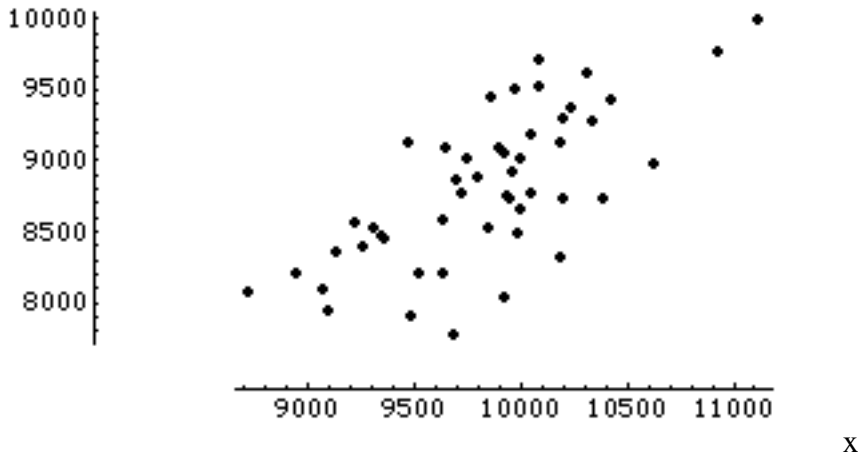
a. Give the usual 95% CI for mu(y) based on the with replacement sample of 50 y-scores.
Ans. yBAR +/- 1.96 s(y) / root(n) = 8810 +/- 1.96 527/root(50) (there are 50 y-scores)
(it works out to [8663.92, 8956.08])

b. Compare the estimate yBAR with the regression-based estimate.
Ans. (yBAR + (mux – xBAR) rhoHAT sy / sx
= 8810 + (9900-9843) 0.694 527/480
(This works out to 8853.43 which is quite different from yBAR = 8810. The fact that xBAR = 9843 < mux = 9900 informed the regression estimator to increase yBAR since x is positively correlated with y).

c. Give the 95% CI for mu(y) based on the regression approach.
Ans. (yBAR + (mux – xBAR) rhoHAT sy / sx +/- 1.96 (sy/root(n)) root(1-rhoHAT^2)
= 8810 + (9900-9843) 0.694 527/480 **+/-** 1.96 (527/root(50)) root(1 – 0.694^2)
(works out to [8748.26, 8958.6]). You can see that it is narrower than (a) as well as being shifted to the right.

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).
Ans. Having seen the regression-based CI (c), which is narrower than (a) by the factor root(1-rhoHAT^2), we see that nFINAL = n / root(1-rhoHAT^2) = 50/(1-0.694^2) = 97 (round up). So we need 97-50 additional samples (all else being equal) for the regular CI (a) to achieve this degree of precision of (c). Those 47 additional samples would have cost us 47 times $400 = $18,800. The total cost of all 97 would have been $38,800. By using the regression approach we achieved that precision for a cost of 50 400 = $20,000.

e. Here is the sample of 50.



The SAMPLE REGRESSION LINE passes through the point (xBAR, yBAR) with slope

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

Calculate this slope and draw the sample regression line in the above plot.
Ans. Plot the point (xBAR, yBAR) = (9843, 8810). The slope is
0.694 527/480 = 0.76.
Let's call the slope 0.75. So to draw the sample regression line in the plot above we put in the point (9843, 8810) then lay off a second point 4 units to the right of it an 3 units up (ration 0.75). Then draw in the line through those two points. What units? Any will do, it is the slope that counts. You could go 500 right and 0.75 500 = 375 up.

f. Since the regression line passes through (xBAR, yBAR) it may be obtained visually by entering xBAR to the horizontal axis and reading off yBAR (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.

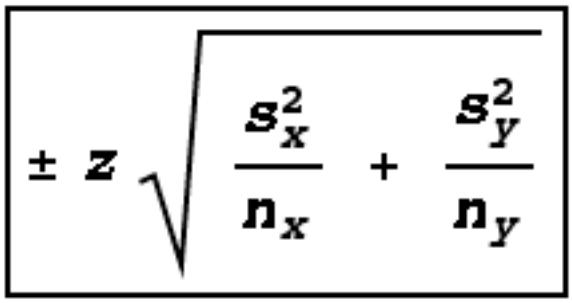
Ans. Enter mux = 9900 to the line you draw in (f). Read off the y. It should be close to the regression estimator 8853.43 (b). It is the very same.

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.
Ans. dBAR = (71 – 25 ... + 50)/16 = 32.81 with s(d) = 55.75 and n = 16
So a 95% z-based CI for mu(current)-mu(previous) is
dBAR +/- 1.96 s(d)/root(n) = 32.81 +/- 1.96 55.75/root(16)
(this is paired data and there are 16 pairs, meaning 16 d-scores). The use of the z-interval is questionable here since 16 is small. If the population of z-scores is normal, or nearly so, then a t-interval with DF = 16-1=15 would be appropriate. In such a case we'd have used t score 2.131.

3. Solve problem 8-2 (paired data).
Ans. We are sampling 40 drivers and reading off x = Mazda time and y = Nissan time for each of them. This is paired data and the times are likely to be dependent (fast drivers have both times low etc.). So to get a CI for mu(d) = mu(x)-mu(y) we use the method of (2) above and process difference scores d = x-y to form the usual 95% z-based CI
dBAR +/- 1.96 s(d)/root(n) = 5 +/- 1.96 2.3 / root(40)

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 n1 and n2 large and the population standard deviations unknown is



Solve 8-22.
Ans. sx = 1.38, sy = 1.56 and xBAR – yBAR = -0.74. This is unpaired data with nx = 15, ny = 23. A 95% z-based CI for mux – muy is
xBAR-yBAR +/- z root(sx^2/nx + sy^2/ny)
-0.74 +/- 1.96 root(1.38^2/15 + 1.56^2/23)

5. Read the Confidence Intervals material beginning at the bottom of pg. 349. Solve problem 8-30 (i.e. give a 95% z-based CI for p1HAT-p2HAT.

Ans. p1 = rate of word of mouth referral small towns, p1HAT = 850/1000 = 0.85

p2 = rate of word of mouth referral metro, p2HAT = 1950/2500 = 0.78

CI for p1 – p2 is

p1HAT-p2HAT +/- 1.96 root(p1HAT q1HAT/n1 + p2HAT q2HAT/n2)

0.85 – 0.78 +/- 1.96 root(0.85 0.15/1000 + 0.78 0.22/2500)

(works out to [0.0425502, 0.0974498], so it seems likely p1 is greater than p2).