

## BONUS Recitation Assignment KEY 2-23-06

This assignment is outside the usual recitation and exam scoring system and will count for up to 4 bonus points towards your course total. Work for clarity, good write-up, accuracy and detail. It is not required that you type your solutions, but you can if you wish. **They must be stapled and every page must have your name on it.**

1. Consult the file POP\_1.pdf file listing data pairs  $(x, y)$  with  $x =$  purchases 2005,  $y =$  purchases 2006 for a large population of customers. In the file an entry of

$\{644, \{924.353, 1253.12\}\}$

means customer #644 has purchases  $x = 924.353$  in 2005 and is projected (by an audit) to purchase  $y = 1253.12$  in 2006.

a. Select a random with-replacement equal-probability sample of  $n = 30$  customers from this population. Take care to explain exactly how you are using Table 14 of random digits to ensure that each customer listed has an equal chance of being drawn into your sample. Do NOT use systematic sampling (e.g. skipping through the file in big steps) or any other method than bona-fide random sampling with the clearest connection to Table 14. There are particular dangers to just grabbing a chunk of 30 from the file since the data of the file is not itself randomly arranged.

Here is MY sample of 30:

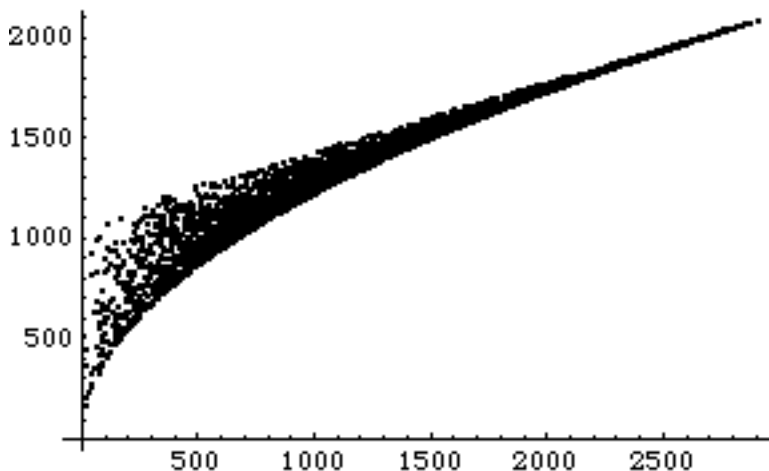
{{848, {781.121, 1089.74}}, {393, {413.987, 802.836}},  
 {768, {1067.81, 1339.48}}, {1859, {1350.8, 1418.45}},  
 {4325, {2769.81, 2031.7}}, {954, {1233.25, 1422.75}},  
 {4523, {1311.99, 1529.73}}, {3189, {2134.45, 1782.54}},  
 {3770, {1017.99, 1367.98}}, {3469, {1852.4, 1673.79}},  
 {3118, {2283.99, 1844.66}}, {391, {652.538, 1083.83}},  
 {2731, {923.814, 1247.25}}, {3853, {2180.92, 1810.25}},  
 {495, {896.985, 1293.57}}, {2572, {2101.85, 1775.26}},  
 {2378, {922.76, 1220.3}}, {1128, {1035.48, 1254.37}},  
 {876, {1007.79, 1265.42}}, {632, {732.591, 1079.93}},  
 {1016, {1015.57, 1250.96}}, {1607, {1474.41, 1496.74}},  
 {3861, {878.538, 1323.}}, {4068, {1064.8, 1408.74}},  
 {152, {507.352, 1154.28}}, {3221, {795.379, 1233.98}},  
 {620, {1015.29, 1346.65}}, {1752, {1437.01, 1468.11}},  
 {4583, {2049.46, 1787.2}}, {2699, {1485.71, 1496.3}}

**Important.** Choose your particular starting point for your use of Table 14 in the following very specific way (so we can check your work if need's be).

**Examine the last two digits of your student number. Choose a column of Table 14 by stepping off the columns in blocks of 4 digits according as the last digit of your student number is 1, 2, through 9. Then choose a row of Table 14 by stepping off blocks of 5 rows according as the second-last digit of your student number is 1 through 9. So a person whose student number ends with "62" will begin with the 30<sup>th</sup> row (6 times 5) and the 8<sup>th</sup> column (2 times 4). They would start with this position in the table then continue either by running through ensuing rows (moving right then**

down to the next row far left), cycling back to the beginning upper left of the table if it is required. The sample of customers produced from these random digits will be used for all of the parts of problem 1.

**FYI.** Here is a plot of all 4612 population (x, y) pairs. As you can see, there appears to be a strong linear relationship but one very disrupted for small x, not an insignificant part of the population. Such a population can induce spectacular failures if samples happen to concentrate, for example, towards the extreme left of the distribution, particularly when regression is used.



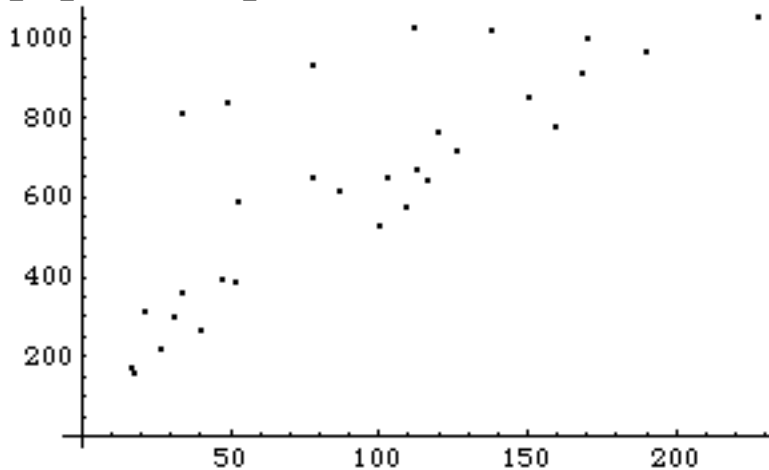
**POP1 means for x, y = {1297.03, 1391.29}.**

**POP1 SDs for x, y = {672.047, 348.926}.**

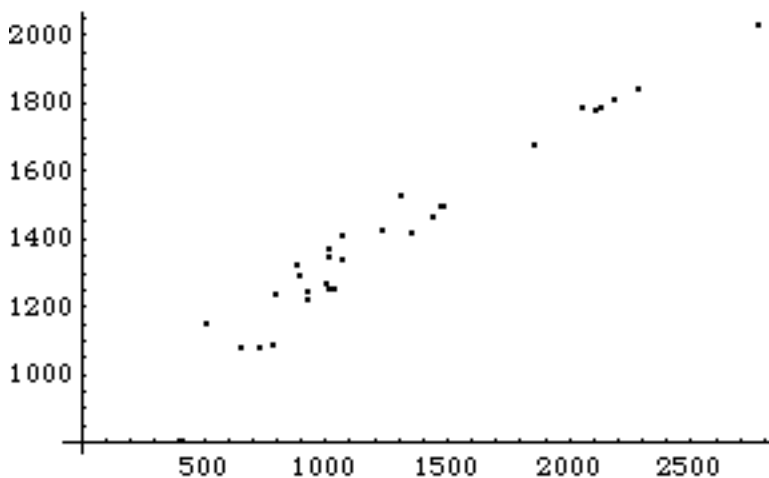
**POP1 correlation rho = 0.967864.**

**You were cautioned against using the first 30 customers instead of a random sample of 30 customers. Sample means for the first 30 customers = {92.249, 637.728} !**

**Here is a plot of the first 30 customers. It is unlike the population plot.**



**Here is a plot of my random sample of 30.**



**For my random sample of 30,**

**$x, y$  means = {1279.86, 1409.99}.**

**$S_x = 587.153, S_y = 275.452.$**

**$\text{rhoHAT} = 0.970545$**

**Compare these with the POP1 values**

**$x, y$  means = {1297.03, 1391.29}.**

**population SDs for  $x, y = \{672.047, 348.926\}.$**

**population correlation  $\text{rho} = 0.967864.$**

b. Estimate the population mean  $\mu_y$  (anticipated purchases for 2006) using the sample mean  $\bar{y}$  and give a 95% CI for  $\mu_y$  based upon this estimator.

**This CI  $\bar{y} \pm 1.96 S_y / \sqrt{30}$  evaluates to  $1409.99 \pm 1.96 \cdot 275.452 / \sqrt{30} = \{1311.42, 1508.56\}$  and does indeed cover  $\mu_y = 1391.29$ , the POP1 mean. Around 5% of students would, due to luck of the draw, miss  $\mu_y$  with their interval.**

c. Re-estimate the population mean for score  $y$  using the regression estimator of  $y$  on  $x$  and give a 95% CI for  $\mu_y$  based upon this regression estimator. To do this you need to know that for this population the mean purchase in 2005 is  $\mu_x = 1297.03$ . Show a plot of the  $(x, y)$  pairs for your sample of 30 customers and how you use it to re-estimate  $\mu_y$  with the aid of this plot and your knowledge that  $\mu_x = 1297.03$  (i.e. how you get the regression estimator).

$$\begin{aligned} \text{regr est} &= \bar{y} + (\mu_x - \bar{x}) \hat{\rho} S_y / S_x \\ &= 1409.99 + (1297.03 - 1279.86) \cdot .970545 \cdot \\ &\quad \cdot 275.452 / 587.153 \\ &= 1417.81 \end{aligned}$$

**Notice that sample mean  $\bar{x} = 1279.86 < \mu_x = 1297.03$ . Thinking I have a low-spending sample I've INCREASED my estimate of  $\mu_y$  from  $\bar{y} = 1409.99$  to  $\text{regr est} = 1417.81$ . Comparing:**

$$\begin{aligned} \bar{y} &= 1409.99 \\ \text{regr est} &= 1417.81 \\ \mu_y &= 1391.29 \end{aligned}$$

**The relevant 95% CI for the regression estimator is**

$$\begin{aligned} & \text{regr est } \pm 1.96 (S_y / \text{root}(30)) \text{root}(1 - \text{rhoHAT}^2) \\ & = 1410 \pm 1.96 (275.452 / \text{root}(30)) \text{root}(1 - .970545^2) \\ & = \{1394.93, 1440.69\} \text{ fails to cover } \text{MU}_y = 1391.29. \end{aligned}$$

**The regression estimator for this data fails utterly. Not only is it actually further from the true value than the ordinary estimate  $\bar{y}$ , but its CI also fails to cover  $\text{MU}_y$  whereas that for  $\bar{y}$  did cover  $\text{MU}_y$ .**

**Other samples will yield different results. If everything is working properly the regression CIs for different samples will be around 24% as wide as the  $\bar{y}$  CIs and each method will cover  $\text{MU}_y$  around 95% of the time.**

d. By what factor is CI (c) narrower than CI (b)? This involves calculating the sample correlation between x and y scores. Show how you do that.

**The CI based upon the regression estimator is narrowed by the factor  $\text{root}(1 - \text{rhoHAT}^2) = 0.241$  (for my sample).**

\*\*\*\*\*

**REPEAT THE ABOVE FOR A 2nd SAMPLE OF 30.**

**{{4443, {2847.41, 2059.94}}, {175, {387.541, 892.948}},  
{3958, {2244.92, 1836.44}}, {575, {946.318, 1301.44}},**

**{2776, {2233.22, 1828.73}}, {2036, {1142.87, 1310.99}},  
{4260, {1168.22, 1460.54}}, {3089, {2155.55, 1791.21}},  
{4274, {2741.74, 2021.32}}, {4608, {2529.46, 1952.68}},  
{2382, {572.257, 1053.57}}, {4120, {2731.44, 2016.63}},  
{1289, {677.042, 1012.99}}, {2295, {1024.48, 1263.83}},  
{1003, {1257.26, 1429.94}}, {2969, {2187.48, 1805.44}},  
{1957, {1205.17, 1341.36}}, {1401, {1422.59, 1482.76}},  
{253, {114.58, 422.021}}, {3535, {1605.03, 1579.09}},  
{2501, {1253.49, 1383.06}}, {2440, {1445.16, 1470.72}},  
{2263, {1222.82, 1358.93}}, {553, {452.69, 823.938}},  
{3278, {2074.15, 1758.56}}, {171, {533.73, 1158.42}},  
{2343, {1970.32, 1721.11}}, {914, {655.11, 987.652}},  
{706, {829.922, 1151.51}}, {2976, {1867.75, 1669.07}}**

**x, y sample means = {1449.99, 1444.89}**

**x, y sample SDs = {776.857, 399.912}**

**sample correlation rhoHAT = 0.97258**

**Compare these with the POP1 values**

**x, y means = {1297.03, 1391.29}.**

**population SDs for x, y = {672.047, 348.926}.**

**population correlation rho = 0.967864.**

**For this data regression is closest to MUy**

**regr est = 1368.31 closest**

**yBAR = 1444.89 furthest**

**MUy = 1391.29 truth**

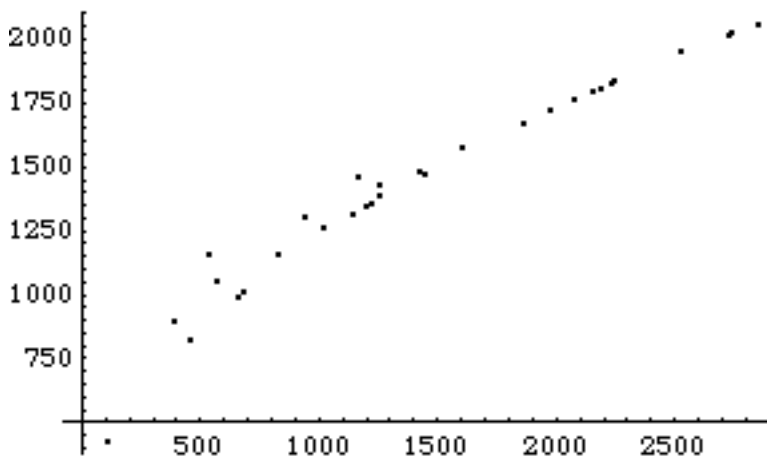
**Here are the CI**

**using yBAR {1301.79, 1588.} wide, covers**

**using regr est {1336.05, 1400.58} narrow, covers**

**For this new data the regr est is doing everything right.**

**Plot of the new sample of 30.**



\*\*\*\*\*

e. Supposing that you had NOT been told  $\mu_x = 1297.03$ , estimate the difference of population means  $\mu_y - \mu_x$  using  $y\text{BAR} - x\text{BAR}$  and give a 95% CI for  $\mu_y - \mu_x$ .

**Hint:** You are simply estimating the population mean  $\mu_{\text{diff}}$  of the difference scores  $\text{diff} = y - x$ . This is called “paired data” (i.e. each customer has two scores  $x, y$ ) and does not require anything new. In particular, don’t use the special method such as would be applicable to independent samples of each of the  $x$  and  $y$  populations.

For my data, the difference scores  $d = y - x$  are:  
{308.615, 388.849, 271.671, 67.6479, -738.113, 189.498,  
217.744, -351.905, 349.987, -178.611, -439.333, 431.29,  
323.437, -370.664, 396.585, -326.59, 297.538, 218.891,  
257.629, 347.338, 235.386, 22.3348, 444.461, 343.939,  
646.925, 438.602, 331.363, 31.1092, -262.258, 10.5848}



**For these scores  $d\bar{B}AR = 130.132$  and  $s_d = 326.627$ . So the PAIRED DATA 95% CI for  $\mu_y - \mu_x$  is**

$$\begin{aligned} & d\bar{B}AR \pm 1.96 s_d / \text{root}(30) \\ & = 130.132 \pm 1.96 326.627 / \text{root}(30) \\ & = \{13.25, 247.014\} \end{aligned}$$

**The true difference of population means is**

$$\mu_y - \mu_x = 1391.29 - 1297.03 = 94.26$$

**lies within the 95% CI for this data.**

f. Take the x scores of your first 15 sample customers and the y scores of your second 15 sample customers. These are now INDEPENDENT samples of 15 from each of two populations (x population and y population). Re-do a 95% CI for  $\mu_y - \mu_x$  using the method for “unpaired independent samples of two populations. It estimates  $\mu_y - \mu_x$  by the difference  $y\bar{B}AR$  (sample of 15) –  $x\bar{B}AR$  (other sample of 15) and the special way of calculating the CI given in the additional slides 2-15-06.

**The CI is**

$$y\bar{B}AR - x\bar{B}AR \pm 1.96 \text{root}( S_x^2 / 15 + S_y^2 / 15 )$$

**For my data this CI is  $\{-385.961, 344.548\}$ . It is far wider than the paired data counterpart in (e) above for this data.**

g. Supposing that you had NOT been told  $\mu_x = 1297.03$  estimate  $\mu_x$  using  $x\bar{B}AR$  and give a 68% (not 95%) CI for  $\mu_x$ . Does your interval contain  $\mu_x = 1297.03$ ? Around 68% of the class should answer yes.

**For my data this is {1172.66, 1387.06}. See (b) for the similar CI for  $MU_y$ . It does contain the population mean  $MU_x = 1297.03$ .**

h. Refer to (g). Has your sample of 30 included any customers more than once? Detail. What is the FPC you are entitled to use for an equal probability without-replacement sample of 30? Would it make any difference to use it?

**There were no repeats in my sample of 30 so I am entitled to view it as a sample without replacement. But what will I gain by doing so? The FPC is**

$$\text{root}((4612-30)/(4612-1)) = 0.99685 \sim 1 \text{ (negl. effect)}$$

i. Use the table of random digits (from your starting point) to produce a with-replacement random sample of 10 listings from the white pages of a Lansing telephone directory for the score  $x =$  first three digits of phone number. Describe exactly what you are doing. Don't phone them. Describe any practical difficulties or "fudges" in the method that you feel you must make in order to carry this off with reasonable effort yet still do a good job of it. Remember, the goal is to avoid a sample that may fall into unfortunate synchrony with the variable  $x$  of interest.

**Suppose there were 893 pages, each with three columns. To produce one sample listing: Sample a page at random (use digit blocks 001 to 893) then one column on that page (use digits blocks 1, 2, 3) then take the next two digits (e.g. 42) as "go 42% down the column." This**

**method is imperfect because not all pages have the same number of listings (nor do columns) and we're "roughing it" when we go (say) 42% of the way down a column. Nonetheless, it is a completely adequate procedure in the practical sense because it is difficult to imagine circumstances in which there could be much danger in using it.**

j. Do you think that just walking up to 30 students on campus at 12 noon and asking them for score  $x = 1$  (on way to lunch) or 0 (not) would yield comparable results to what would be learned from randomly sampling 30 students beforehand and shadowing them and "pouncing on them" at noon? Why?

**People on their way to lunch are out "on their way." We may catch some walkers on our "walk about" but probably not ones comprising a random sample in which each target person has the same chance of inclusion (e.g. lunch could be over before we get far along so we've missed everyone but those on our little trajectory). In particular, anyone driving will be missed altogether, etc. The "identify and pounce" method is much more likely to produce an accurate result but of course it is impractical. Perhaps we could interview those identified as pounce-sample persons the next day?**

k. Refer to (j). What if the score were  $y =$  last digit of student number? Would the two sampling methods produce similar results? Why?

**With last digit of student number it is hard to see how we could get a less than random sample no matter what we do. It is almost as if each person we encounter were assigned a digit at random 0 through 9, independently. True, the digits are really not random, being assigned serially (one would suppose) at some point in the admitting process.**

**Conspiracy theorists can imagine that some of us are assigned last digit "2" (say) because we frequently walk to our lunch. If so, our walk about sample method is biased.**