STT 315
Slides for week 7, 2-20-06 (see also Additional Slides for Week 6)

These slides cover

A. Use of random digits.
B. Regression estimator.

A. Use of random digits. This topic was introduced early in the course but will be more fully elaborated here since it is needed for the BONUS assignment due this week in recitation.

Random digits, such as found in Table 14 of your textbook, appear something like this:

Table 14 Random Numbers

1559	9068	9290	8303	8508	8954	1051	6677	....
5550	6245	7313	0117	7652	5069	6354	7668	....

The idea is that these digits should behave as though they were produced from with-replacement and equal-probability sampling of the ten digits {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}. For example, any particular digit, such as "7" would be expected to occur in roughly one of ten times. Likewise, consecutive pairs 15 59 90 68 .. can be regarded as with-replacement samples of the 100 two digit pairs {00, 01, ..., 99}. The independence of such digits means that if portions of the table are revealed to us the odds are not changed for the portions unseen. As an example, upon seeing the first block 1559, the conditional probability that the very next block is also 1559 is one in 10000, just as it would be if you had not seen the first block of four.

How do we use random digits to effect a random sample? Basically, we set up a 1:1 corresspondence between the population units and random digit patterns so that everybody gets the same chance to be chosen. For example, if your population has 53996 units and a with-replacement sample of 4 units is desired you could set up the corresspondence

unit 1 ⟷ digit pattern 00001
unit 2 ⟷ digit pattern 00002
... ...
unit 53996 ⟷ digit pattern 53996

Using the portion of Table 14 above we can decide to take consecutive non-overlapping blocks of 5 digits, skipping any greater than 53996 and the 00000 block (skipping them does not alter the odds for those actually used). Here is the sample we obtain by this method:

Table 14 Random Numbers

1559 9	068 92	90 830	3 8508	8954 1	051 66	77 ....

so units 15599, 6892, 38505, 5166 comprise our with-replacement sample of four. If we desire a sample without-replacement it is only necessary to skip over any five digit block that has previously occurred, again not changing the odds for ones that are selectable.

B. Regression estimator. The idea of drawing a straight line through a cloud of (x, y) points is very old and has many applications. We'll describe a way to use a line through points to narrow a confidence interval. For example, suppose we wonder how much revenue will come to us this year from a population of ten thousand rental properties, each of which is subject to its own local economy with differing tax, economic health, maintainance, and other issues. We have recourse to sampling (say) 100 of the properties and audit-

ing them to learn (maybe predict as best we can) how much revenue each will produce. If would be enormously costly to do this for all ten thousand properties. Let's focus on the mean revenue y per property. That can project total revenue by multiplying the CI by ten thousand. The usual 95% CI for μ_y is

$$\text{yBAR} \pm 1.96 \frac{s_y}{\sqrt{100}}$$

We can narrow this 95% interval by increasing the sample size n = 100 but this comes at some cost. Perhaps it costs $800 to audit each sample property. To double precision will require around n = 400 sample size which adds 300 times $800 = $240,000 to the cost! What if we could effectively narrow the interval without much additional cost?

Here is a way to do it using something called a **regression estimator**. What you do is score each sample property with (x, y) where y is the projected revenue and x is what we earned from that property last year (a matter of record). It will turn out to be advantageous for us to record x for each sample property. This comes at virtually no additional cost and will narrow the CI for the same n = 100 sampling effort (and cost).

After our sample of 100 properties is in we calculate the following five (so-called first and second order) statistics:

$\overline{x}$	$\overline{y}$	
$\overline{x^2}$	$\overline{y^2}$	$\overline{\text{xy}}$

We surely know the average revenue μ_x of all ten thousand properties from last year. Also, it seems reasonable to suppose that there is some degree of **positive linear association** between revenues last year and this (i.e. a plot of all ten thousand (x, y) scores, if it could be had, would likely show a cloud of points around an upward sloping line, since higher than average revenue x last year is likely to be accompanied by higher than average revenue y this year, and likewise low x will be associated with low y, not perfectly by approximately).

The basic idea. If our sample 100 properties has $\bar{x} < \mu_x$ we reason that $\bar{y}$ is also likely to be lower than μ_x. So we might improve upon the estimate by increasing $\bar{y}$ in such a case according to how far below μ_x our sample $\bar{x}$ has fallen and the apparent degree of linear association between x and y revealed by our sample of 100 properties.

The regression estimator adjusts $\bar{y}$ as follows:

$$\text{regression estimator } \hat{\mu}_{y,\,\text{regr}} = \bar{y} + (\mu_x - \bar{x})\,\hat{\rho}\,\frac{\sqrt{\overline{y^2} - (\bar{y})^2}}{\sqrt{\overline{x^2} - (\bar{x})^2}}$$

where

$$\text{sample correlation } \hat{\rho} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\,\sqrt{\overline{y^2} - (\bar{y})^2}}$$

What is the payoff for using this regression estimator? It can be seen in the form of the 95% (or other) CI for μ_y.

$$\bar{y} \pm 1.96 \frac{s_y}{\sqrt{100}}$$ ignoring x

$$\hat{\mu}_{y,\,\text{regr}} \pm 1.96 \frac{s_y}{\sqrt{100}} \sqrt{1 - \hat{\rho}^2}$$ using x

Since $-1 \leq \hat{\rho} \leq 1$ the shrinkage factor is $0 \leq \sqrt{1 - \hat{\rho}^2} \leq 1$. If the sample (x, y) points fall exactly on a straight line of (some) upward slope $\hat{\rho} = 1$ and this shrinkage factor is $0 = \sqrt{1 - 1^2}$ indicating perfect prediction. In effect, you know μ_y for this year since you know μ_y from last year and there appears to be perfect positive correlation. The same would be true if there is perfect negative correlation $\hat{\rho} = -1$, i.e. all points on a line of (some) downward slope.

The above is all you need for your assignment due 2-23-06. Read on for general information about random sampling and also regression.

A. More about uses of random digits.

How are the digits of Table 14 produced? One might imagine a scheme in which a sort of roulette wheel with sectors {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} is spun repeatedly. Rand Corporation used an electronic roulette wheel to produce digits for their book "A **Million Random Digits** with 100000 Normal Deviates" The Free Press, 1955. Such efforts go back at least to 1927. Other mechanisms designed to tap physical randomness abound, one current example being a very high speed electronic device (about the size of a small refrigerator) to monitor electronic