# You should know...

- how to choose a random number
- auxiliary variables and ratio estimation of $\bar{y}_U$
- new notation $t_x$, $t_y$, $B$, $R$, $S_{xy}$, $S_x^2$, $S_y^2$
- and $\widehat{\bar{y}}_r$, $\widehat{B}$, $\widehat{t_{yr}}$, $\widehat{t}_x$
- TEST: Chapter 1, Chapter 2 EXCEPT §2.8, Chapter 3.1, lectures

  Friday, October 16, 1.10 to 2 pm, SS 2117

  1 side 8 1/2 x 11 " sheet of notes

  1 non-programmable calculator

  sit in every 2nd seat

# Using a random number table

- HW: Exercise 2.4
- a: N = 742; n = 30; take sequences of 3 digits, ignore any repeats or numbers $>$ 742 ✓
- b: if number $>$ 742 eliminate first digit, start sequence with next digit ✓
- c: N=170; use the rounded remainder ✗thanks to someone for pointing out that 5*170 = 850, so for the group of digits between 850 and 999 you get only numbers between 1 and 149
- d: N=200; take 2 digit numbers, put a decimal place, .75, .43, .01, etc. and multiply by 200 ✗(no odd numbers)
- e: school of 20 classes; each class has 20-40 students; select a class at random and a student within a class at random ✗
- f: school again: choose pairs of random numbers

# ... random numbers

- HW: Exercise 2.24
- p.414: Stephens County; Lockhart city: districts 51-75, each district has between about 500 and 1300 houses (p.416)
- a: randomly select a district between 51 and 75
- b: randomly select a house from those in the chosen district
- c: reject if the house is already in the sample
- d: repeat until sample size is $n$
- is this SRS? why not? (Exercise 2.25 works)

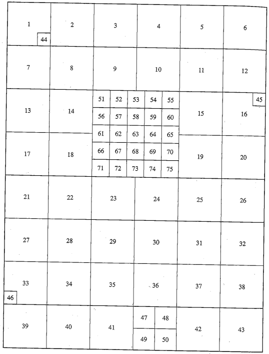



**FIGURE A.1**
A district map of Stephens County

| Area | Districts | Number of houses |
|---|---|---|
| Rural areas | 1–43 | 7,932 |
| Lockhart City | 51–75 | 19,664 |
| Eavesville | 47–50 | 3,236 |
| Villegas | 44 | 283 |
| Weldon | 45 | 562 |
| Routledge | 46 | 312 |

# Ratio Estimation

- why?
- want to estimate a ratio, e.g. average yield per acres, percentage of magazine pages devoted to advertising, ...
- want to estimate a population total, but $N$ is unknown (Ch. 12.2)
- increase precision of estimate of $t_y$ or $\bar{y}_U$
- adjust estimates from a sample to reflect demographics (p.62) – post-stratification (Ch. 4, 7, 8)
- adjust for non-response (Ch. 8)

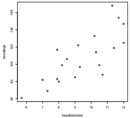
# ... ratio estimation

- $$\widehat{t}_{yr} = \widehat{B} t_x = \hat{t}_y \left( \frac{t_x}{\hat{t}_x} \right)$$
- $$\widehat{\bar{y}}_r = \widehat{B} \bar{x}_U = \bar{y} \left( \frac{\bar{x}_U}{\bar{x}} \right)$$
- $$\widehat{B} = \frac{\bar{y}}{\bar{x}}$$
- $$E(\widehat{\bar{y}}_r) \approx \bar{y}_U, \quad \hat{V}(\widehat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n}$$
- $$E(\widehat{t}_{yr}) \approx t_y, \quad \hat{V}(\widehat{t}_{yr}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n}$$
- $$e_i = y_i - \widehat{B} x_i, i = 1, \ldots n$$

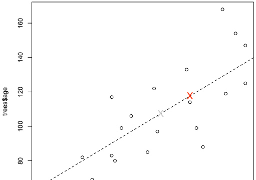
# Exercise 3.1

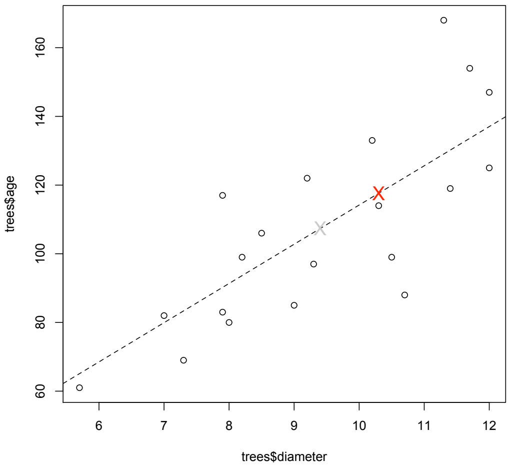
- a: estimate the proportion of time devoted to sports in television news broadcasts
- $x_i$ – length of news broadcast $i$, $y_i$ – length of sports coverage in n.b. $i$ (Good Housekeeping, p.61)
- b: estimate the average number of fish caught per hour by anglers visiting a lake
- $x$ – number of hours fished, $y$ – number of fish caught (GH, p.61)
- c: estimate the average amount that undergraduate students spent on textbooks
- $x$ – number of textbooks bought, $y$ – total cost
- d: estimate the total weight of usable meat in a shipment of chickens
- $t_x$ – total weight of shipment; $(x_i, y_i), i \in \mathcal{S}$ – weight of chicken $i$, amount of usable meat in $i$

# Exercise 3.4



- a: plot the data
- b: $\bar{x}_U = 10.3$, $\bar{x} = 9.405$, $\bar{y} = 107.4$, $\widehat{B} = 11.41946$
- $\widehat{\bar{y}}_r = \widehat{B}\bar{x}_U = 117.6$
- $$\hat{V}(\widehat{\bar{y}}_r) = \left(1 - \frac{20}{1132}\right)\frac{321.933}{20} = 16.09665 = (3.98)^2$$
- 95% confidence interval for $\bar{y}_U$ : $117.6 \pm 1.96 \times 3.98$
- ignoring $x$ get $117.6 \pm 1.96 \times 6.35$





160
140
120
100
80
60
trees$age
6
7
8
9
10
11
12
trees$diameter

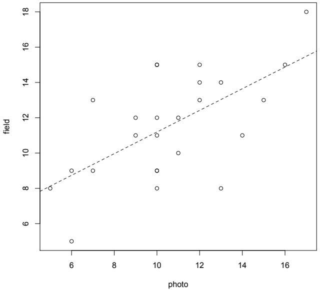
# Understanding §3.1

- 3.1.2: derivation of bias and MSE – single trick: replace $\bar{x} = \bar{x}_{\mathcal{S}}$ in denominator by $\bar{x}_U$
- wave hands to say bias $\approx 0$ hence $V(\widehat{B}) \approx E(\widehat{B} - B)^2$
- leads to Equations (3.8) and (3.9)
- some rough calculations pp. 69,70 to show: $\approx$ is okay if $n > 30$ and $CV(\bar{x}) \leq 0.1$, $CV(\bar{y}) \leq 0.1$
- 3.1.2.2: it pays to use ratio estimation if $\mathrm{corr}(x, y) > 0.5$
- 3.1.3: use the same formula with proportions; here there is no $N$, so ignore $(1 - n/N)$

# Regression estimation (not on test)

- instead of
$$\widehat{\bar{y}}_r = \widehat{B}\bar{x}_U$$
-
$$\widehat{\bar{y}}_{reg} = \widehat{B}_0 + \widehat{B}_1\bar{x}$$
-
$$\widehat{B}_1 = \frac{\sum_{i\in\mathcal{S}}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i\in\mathcal{S}}(x_i - \bar{x})^2}$$
-
$$\widehat{B}_0 = \bar{y} - \widehat{B}_1\bar{x}$$
-
$$SE(\widehat{\bar{y}}_{reg}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}$$
- where now
$$s_e^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \widehat{B}_0 - \widehat{B}_1 x_i)^2$$

# Example 3.6




- $\bar{x}_U = 11.3$
- $\widehat{B}_0 = 5.06$, $\widehat{B}_1 = 0.6133$
- $\widehat{\bar{y}}_{reg} = 5.06 + 0.613(11.3) = 11.99$
- $SE(\widehat{\bar{y}}_{reg}) = \sqrt{(1 - 25/100)5.54834/25} = 0.408$ improvement over $\bar{y}$
- End of Chapter 3