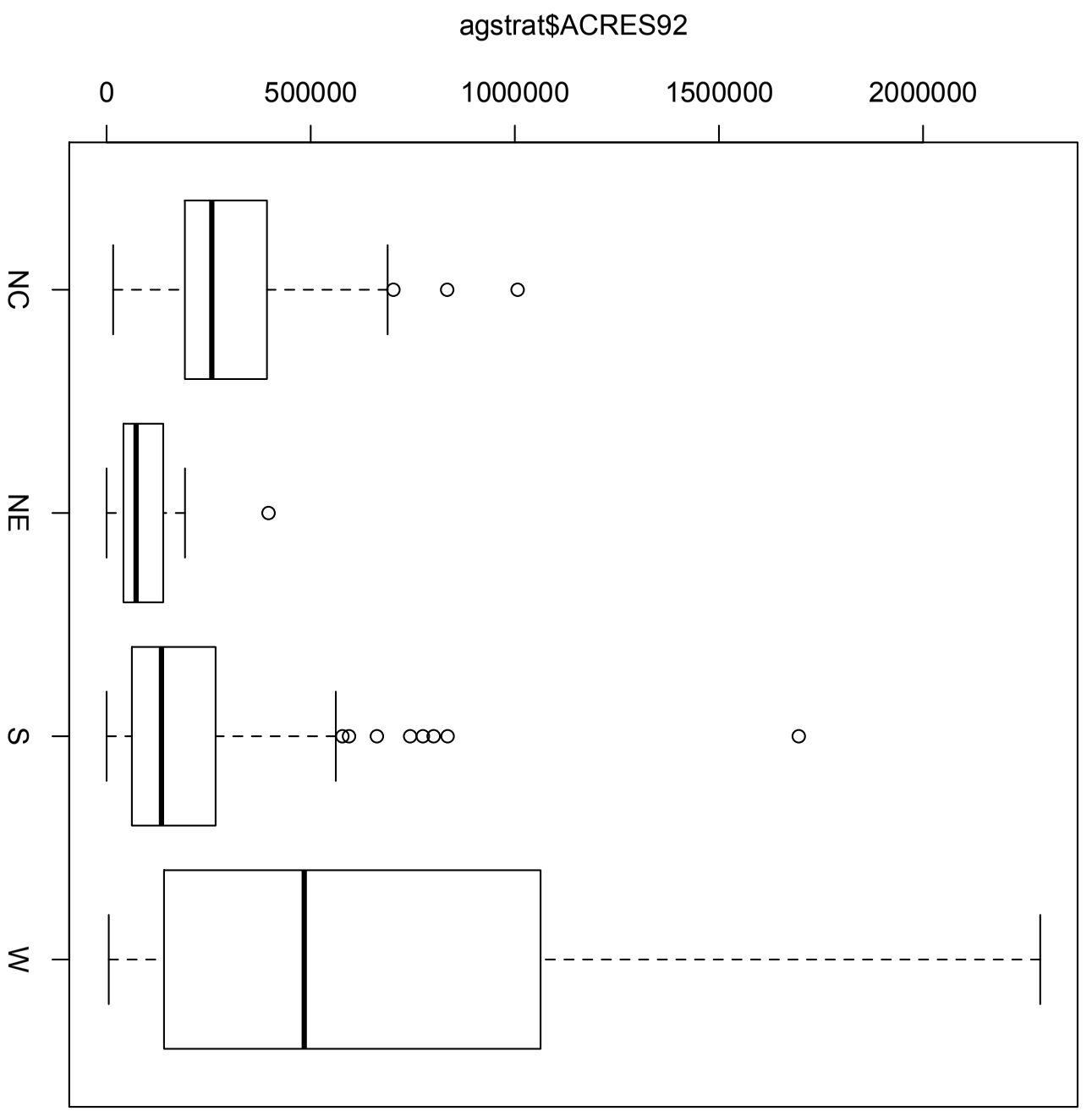
You should know...

- regression estimation §3.2
- $\widehat{\bar{y}}_{reg} = \bar{y} + \widehat{B}_1(\bar{x} - \bar{x}_U) = \widehat{B}_0 + \widehat{B}_1\bar{x}$
- $\hat{V}(\widehat{\bar{y}}_{reg}) = \left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}$
- $s_e^2 = \sum_i (y_i - \widehat{B}_0 - \widehat{B}_1 x_i)^2/(n-1)$
- HW: Exercise 3.5, 3.13a, 3.15*, 3.24
- For interest: §3.2.2; difference estimation in auditing, §3.3; ratio estimation in domains
- Omit §3.4
- ratio estimation: y is correlated with x, $V(y)$ increases with x, $y = 0$ when $x = 0$
- regression estimation: y is correlated with x, $V(y)$ constant in x, $y \neq 0$ when $x = 0$
- **Final Exam: December 15, 9 - 11 a.m., EX 200 (255 McCaul Street)**
- $$: Samuel Beatty Scholarship November 13

http://www.physics.utoronto.ca/undergraduate/scholarship.2009

Stratified sampling Ch.4 

- divide the population into strata: ...
- choose a sample from each strata, using a probability method ...
- offers some protection that the sample is representative ...
- want particular precision for subgroups ...
- may be cheaper
- may be more precise ... (variance within)
- Example 4.1: stratify farms in US by region: Northeast, North Central, South, West


agstrat$ACRES92
0
500000
1000000
1500000
2000000
NC
NE
S
W

Estimation 

#	Stratum	Sample Size	Average	Variance	Estimated Total of Farm Acres	Estimated Variance of Total
220	Northeast	21	97,629.8	7,647,472,708	21,478,558	1.59432×10^{13}
1054	North Central	103	300,504.2	29,618,183,543	316,731,379	2.88232×10^{14}
1382	South	135	211,315.0	53,587,487,856	292,037,391	6.84076×10^{14}
422	West	41	662,295.5	396,185,950,266	279,488,706	1.55365×10^{15}
	Total	300			909,736,034	2.54190×10^{15}

each line is calculated using theory of SRS

e.g. NE: $\bar{y} = 97,629.8$; $N = 220$; $\hat{t} = N\bar{y}$

$$V(\hat{t}) = (220)^2(1 - \frac{21}{220})\frac{s^2}{21} = 1.5943 \times 10^{13}$$

Notation 

#	Stratum	Sample Size	Average	Variance	Estimated Total of Farm Acres	Estimated Variance of Total
220	Northeast	21	97,629.8	7,647,472,708	21,478,558	1.59432×10^{13}
1054	North Central	103	300,504.2	29,618,183,543	316,731,379	2.88232×10^{14}
1382	South	135	211,315.0	53,587,487,856	292,037,391	6.84076×10^{14}
422	West	41	662,295.5	396,185,950,266	279488,706	1.55365×10^{15}
	Total	300			909,736,034	2.54190×10^{15}

$N_h \quad h \qquad\qquad n_h \quad \bar{y}_h \quad s_h^2 \qquad \hat{t}_h = N_h \bar{y}_h \qquad \hat{V}(\hat{t}_h)$

$$s_h^2 = \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2/(n_h - 1), \qquad V(\hat{t}_h) = N_h^2(1 - n_h/N_h)s_h^2/n_h$$

every line just uses SRS theory; see p.99

Now look at total

Estimating $\bar{y}_U$ and t (p.100)

Total	300	909,736,034	2.54190×10^{15}

$$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h$$

$$\bar{y}_{str} = \frac{\hat{t}_{str}}{N}$$

$$V(\hat{t}_{str}) = \sum_{h=1}^{H} V(\hat{t}_h) =$$

$$\widehat{V}(\hat{t}_{str}) =$$

$$\widehat{V}(\bar{y}_{str}) = \widehat{V}(\hat{t}_{str})/N^2 =$$

Practise

- HW: Example 4.2, Example 4.3, Exercise 4.9.2
-

```
> ex42
  stratum  Nh nh ybarh    s2h
1       A 400 98  24.1   5575
2       B  30 10  25.6   4064
3       C  61 37 267.6 347556
4       D  18  6 179.0  22798
5       E  70 39 293.7 123578
6       F 120 21  33.2   9795
> ex42$thath = ex42$Nh*ex42$ybarh
> ex42$Vhath = (1-ex42$nh/ex42$Nh)*ex42$Nh^2*ex42$s2h/ex42$nh
> sum(ex42$thath)
[1] 54496.6
> sum(ex42$Vhath)
[1] 34105732
> sqrt(34105732)
[1] 5840.011
> 54497 + 1.96*5840.011
[1] 65943.42
> 54497 - 1.96*5840.011
[1] 43050.58
```

Another look at the formulas 

- $$\hat{t}_{str} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}$$
- $$w_{hj} = \frac{1}{\pi_{hj}} = \frac{N_h}{n_h}$$
- $$\bar{y}_{str} = \frac{\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj}}$$
- w_{hj} Sampling Weights $w_{hj} = w_h = N_h/n_h$
- how many units in the population does one observation "stand for"?
- agriculture: $N_h = 220, 1054, 1382, 422$; $n_h = 21, 103, 135, 41$, $w_h = 10.2, 10.5, 10.2, 10.3$

How many observations in each stratum?

- proportional allocation: $\pi_{hj} = n/N$ (same for each h)
- Example: 10% sample of 100 women and 200 men: $n_{women} =$, $n_{men} =$
- sample is called self weighting: $w_h = N/n$
- can show that $V(\bar{y}_{str})$ usually smaller than $V(\bar{y})$, especially if the stratum means are quite different (Table 4.3)

... how many observations

- optimal allocation: $n_h \propto \frac{N_h S_h}{\sqrt{c_h}}$
- c_h the cost of taking an observation in stratum h
- $S_h^2 =?$; do we know this? what to do?
  see Example 4.8
- Example 4.6 ASCAP catalogue
  - sampling from producers' cue sheets
  - identify musical compositions
  - allocate royalties by estimating proportions for each composer
  - strata are based on: size of license fee paid by radio station, geographic region, type of community
- Example 4.7: Dollar stratification in accounting
  - loans more than $ 1 million
  - loans between $500,000 and $ 1 million
  - etc. down to loans under $ 10,000
  - S_h^2 larger in the higher loan strata
- HW: Exercise 4.9.12

... how many observations

- Now we know the relative sample sizes, but what about the total sample size $n = \sum n_h$?
- §4.4.4: use our old formulas and ignore fpc
- which is better: proportional or optimal?
- it depends...
- most surveys record more than one variable

How do we define strata? §4.5

- need strata to be of particular interest
- or to be very efficient
- efficient means large between strata variability
- i.e. stratum means $\bar{y}_{hU}$ very different
- why would we ever use SRS? easier, cheaper, gain in efficiency might not be worth it...
- example: Canadian Survey of Employment, Payroll and Hours: industry/province/estimated payroll
- example: Nielsen ratings: geographic region, county size, amount of cable service, ...
- stratification adds to the cost and complexity
- HW: read Example 4.10

In the News

Afghanistan may yet avoid run-off

Afghan President cedes to Western pressure in announcing second vote, but many hopea power-sharing agreement can be reached first – and prevent more election violence

Quota Sampling