Comparing Stratified to SRS

Week 3 (3.4)

Stat 260, St. Clair

When is a Stratified sample more precise than SRS?

When does

$$SE(\hat{t}_{str}) \overset{???}{<} SE(\hat{t}_{SRS})$$

$$SE(\bar{y}_{str}) < SE(\bar{y})$$

answer: It depends on the measurement's **Analysis of Variance** (ANOVA)

Lohr Examples 3.2 and 3.6: Design effect

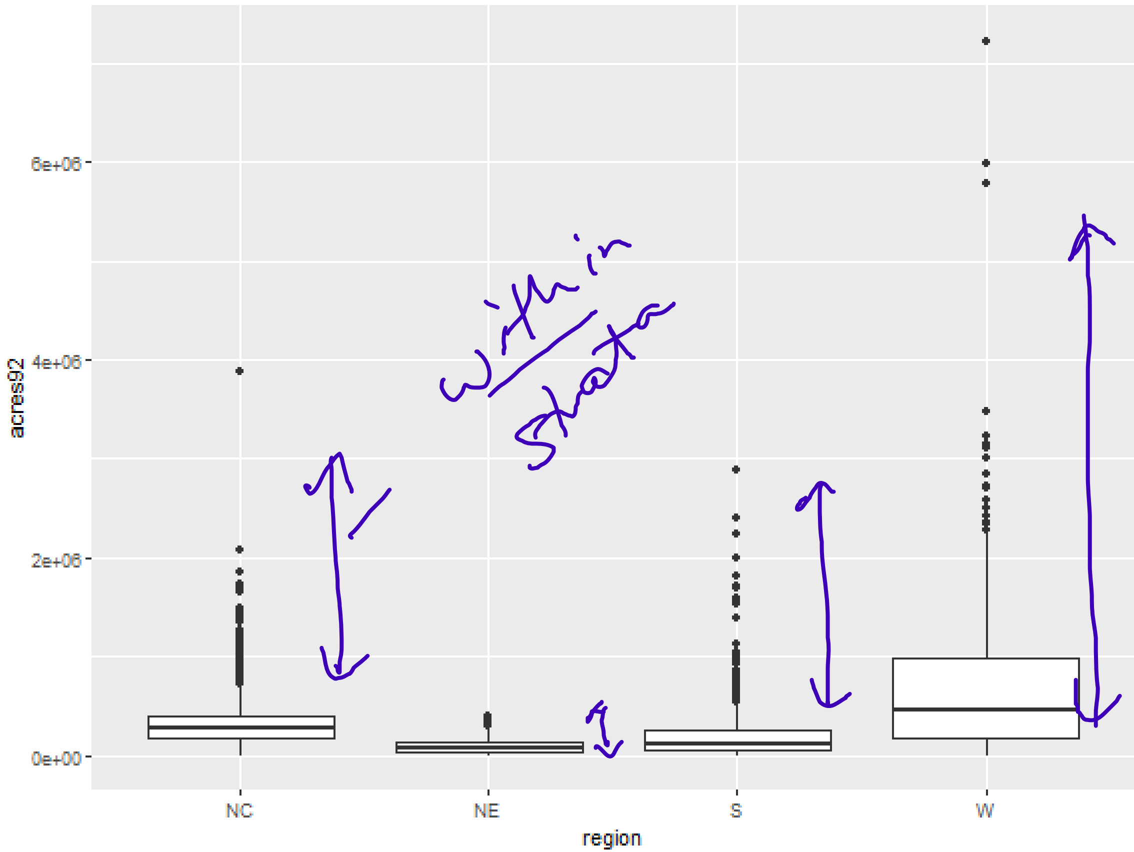
```
> # Design effect comparing stratified to SRS
> svytotal(~acres92 + farms92, design_strat, deff=T)
             total         SE   DEff
acres92 909736033  50417248 0.7945
farms92   1961190     74726 0.9751
```

- acres92: stratified design is more precise that SRS
- farms92: about same

Lohr Examples 3.2 and 3.6: `acres92` by strata

```
> ggplot(agpop, aes(x=region, y = acres92)) +
+   geom_boxplot()
```

overall variation
total

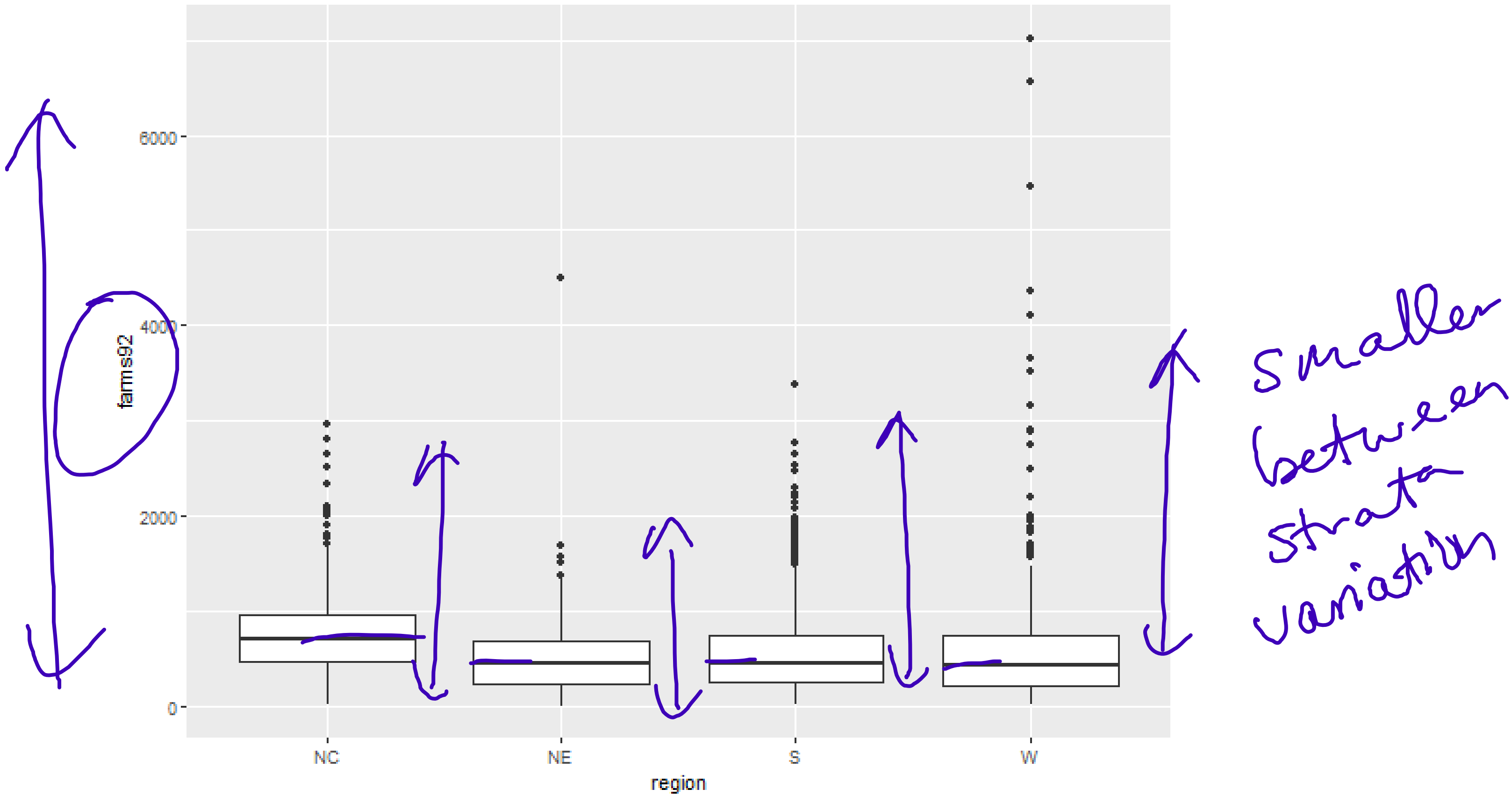


within strata

between strata
compares strata means

Lohr Examples 3.2 and 3.6: `farms92` by strata

```
> ggplot(agpop, aes(x=region, y = farms92)) +
+   geom_boxplot()
```



Population ANOVA

Let y_{hj} be your measurement.

ANOVA breaks the **total** sum of squares of y into **between strata** and **within strata** variation:

$$\boxed{SST = SSB + SSW}$$

S^2 = pop. variance of y's.

- $SST = \sum_{h=1}^{H} \sum_{j=1}^{N_h} (y_{hj} - \bar{y}_{\mathcal{U}})^2 = (N-1)S^2$

overall pop. mean

- $SSB = \sum_{h=1}^{H} N_h (\bar{y}_{h,\mathcal{U}} - \bar{y}_{\mathcal{U}})^2$

stratum h pop. mean

S_h^2 = var. of y's in str. h (pop.)

- $SSW = \sum_{h=1}^{H} \sum_{j=1}^{N_h} (y_{hj} - \bar{y}_{h,\mathcal{U}})^2 = \sum_{h=1}^{H} (N_h - 1) S_h^2$

Variance: SRS

For a SRS of size n, we can write the variance, SE^2, of $\hat{t}_{SRS}$ as

$$Var(\hat{t}_{SRS}) = N^2 \left(1 - \frac{n}{N}\right) \frac{SSB + SSW}{n(N-1)}$$

where S is the SD of the measurements in the population.

proof

$$Var(\hat{t}_{SRS}) = N^2\left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

$S^2 \to$ pop. var.

$$S^2 = \frac{SST}{N-1}$$

Variance: Stratified sample

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

For a stratified sample, assume

- overall sample size is $n = n_1 + \cdots + n_h$
- we used **proportional allocation** to determine stratum sample sizes:

$$n_h = n \times \frac{N_h}{N}$$

$$\underbrace{\frac{n_h}{n}}_{\text{fraction of sample in str. } h} = \underbrace{\frac{N_h}{N}}_{\text{fraction of pop in str. } h}$$

Variance: Stratified sample

For a stratified sample with proportional allocation, we can write the variance of $\hat{t}_{str}$ as

$$Var(\hat{t}_{str}) = N\left(1 - \frac{n}{N}\right)\frac{\sum_{h=1}^{H} S_h^2 + SSW}{n}$$

where S is the SD of the measurements in the population.

proof

$$V(\hat{t}_{str}) = \sum_h \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h} = \text{algebra}$$

$\Rightarrow$ plugin $n_h = n\left(\frac{N_h}{N}\right)$

$$\Rightarrow SSW = \sum_h (N_h - 1) S_h^2 = \underbrace{\sum_h N_h S_h^2} - \sum_h S_h^2$$

Variance: SRS vs. Stratified sample

plugin $Var(\hat{t}_{SRS})$
$Var(\hat{t}_{str})$
& algebra

Using proportional allocation,

$$SE(\hat{t}_{str}) < SE(\hat{t}_{SRS})$$

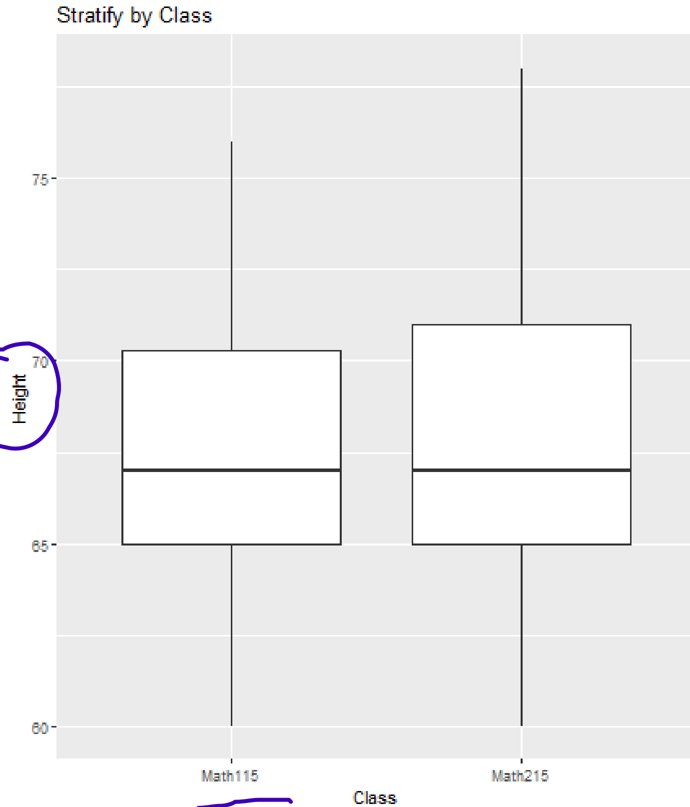
when

$$\boxed{\sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) S_h^2 < SSB}$$

Strat (prop alloc) better than SRS

- Large between strata differences (SSB)
  $\bar{y}_h$ very different!
- More homogeneous responses within strata (compared to overall variation)
  S_h^2 small!

Example: One Population, two stratifications



Strata = Class: $SSB = 0.0007$

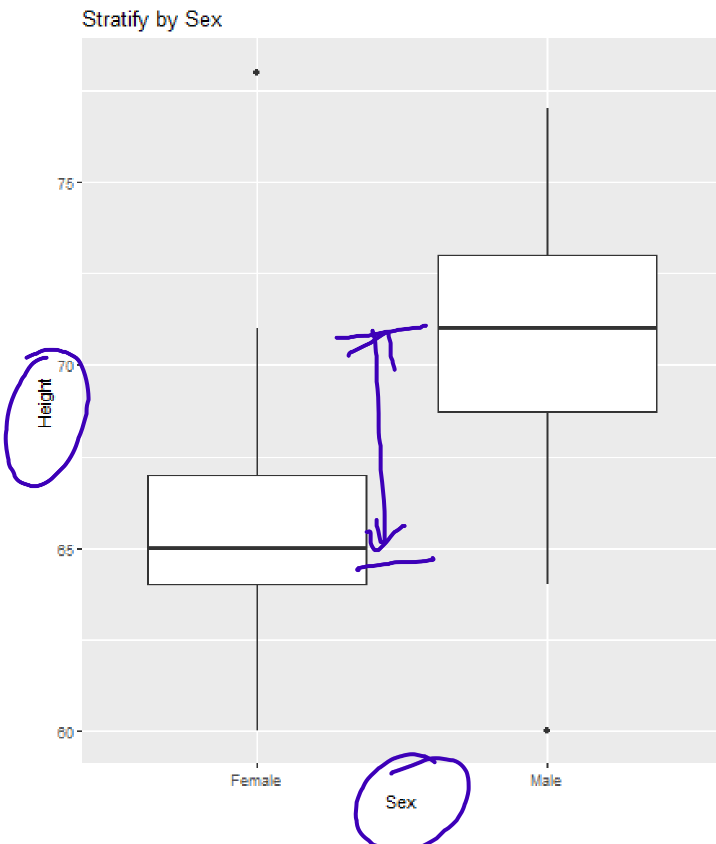
- 115 population: $N_{115} = 71$, $S^2_{115} = 15.7$
- 215 population: $N_{215} = 57$, $S^2_{215} = 17.8$

$$\sum \left(1 - \frac{N_h}{N}\right) S_h^2$$

$$\left(1 - \frac{71}{128}\right) 15.7 + \left(1 - \frac{57}{128}\right) 17.8$$

$$\approx 16.9 \quad \cancel{\approx} \quad .0007$$

Example: One Population, two stratifications



Strata?

Strata = Sex: $SSB = 846.0$

- Female population: $N_F = 68$, $S_F^2 = 8.4$
- Male population: $N_M = 60$, $S_M^2 = 11.7$

$$\left(1-\frac{68}{128}\right)8.4+\left(1-\frac{60}{128}\right)11.7$$

$$=10.2 < 846$$

Strata based on Sex yield more precise est. than SRS

Post-Hoc comparison

Q: How do we compute the design effect with a **sample of data** from any stratified sample? (not just one with proportional allocation)

$$DEff = \frac{V(\bar{y}_{str})}{V(\bar{y}_{SRS})} = \frac{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}}{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}}$$

- $V(\bar{y}_{str})$: estimate from the SE of your stratified data
- $V(\bar{y}_{srs})$: need to use the **stratified** sample to get an unbiased estimate of population measurement's variance S^2

if SRS data : $\hat{S}^2 = s^2$ (sample var.)

if Strat data : s^2 (sample var) biased est. of S^2

Post-Hoc comparison

$V(\bar{y}_{srs})$: need to use the **stratified** sample to get an unbiased estimate of population measurement's variance S^2

1. Use sampling weights (example 7.4) to estimate S^2

$$\hat{S}^2 = \frac{N}{N-1}\left[\sum y_i^2\left(\frac{w_i}{\sum w_j}\right) - \left(\sum y_i\left(\frac{w_i}{\sum w_j}\right)\right)^2\right]$$

* survey package

Post-Hoc comparison

$V(\bar{y}_{srs})$: need to use the **stratified** sample to get an unbiased estimate of population measurement's variance S^2

2. Estimate the population **sum of squares** values from your stratified data's ANOVA:

$$\hat{S}^2 = \frac{\widehat{SST}}{N-1} = \frac{\widehat{SSB} + \widehat{SSW}}{N-1} \quad \to \text{pop.}$$

where SS are estimated from the stratified data as

$$\widehat{SSW} = (N-H)msw_{sample} \qquad \widehat{SSB} = \sum_{h=1}^{H} N_h(\bar{y}_h - \bar{y}_{str})^2$$

average within SS in sample

$$SSB = \sum_h N_h(\bar{y}_{hU} - \bar{y}_U)^2$$

Lohr Examples 3.2 and 3.6: Design effect

Compare stratified to SRS when estimating the mean number of large farms (`largef92`) in 1992 in the US:

```
> # Design effect comparing stratified to SRS
> svymean(~largef92, design_strat, deff=T)
            mean      SE   DEff
largef92 56.6980  3.5577 0.865
```

Let's estimate this DEff "by hand"

- Numerator is 3.5577^2
- Denominator: need to estimate S^2 (SD of `largef92` in the **population**)

Lohr Examples 3.2 and 3.6: Design effect

Here we model `largef92` as a function of `region` (**strata**) and use `anova` to get the **sample anova table**:

y ~ strat

```
> largef92_lm <- lm(largef92 ~ region, data=agstrat)
> # "Residuals" == WITHIN strata
> # region (strata) == BETWEEN strata
> anova(largef92_lm)
Analysis of Variance Table

Response: largef92
           Df  Sum Sq Mean Sq F value    Pr(>F)
region      3  208850   69617  16.555 5.699e-10 ***
Residuals 296 1244705    4205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

strata

within

$$msw_{sample} = 4205$$

$$\widehat{SSW} = (\underset{N}{3078} - \underset{H}{4})\,4205 = \boxed{12,926,170}$$

Lohr Examples 3.2 and 3.6: Design effect

Between Strata sum of squares:

- Overall estimated mean from stratifed estimator:

```
> ybar_str <- 56.6980
```

$\bar{y}_{str}$

- Within strata mean estimates:

```
> mn_region <- agstrat %>%
+                   group_by(region) %>%
+                   summarize(mean(largef92)) %>%
+                   pull()
> mn_region
[1]  70.912621   8.190476  38.837037 104.975610
```

NC NE S W

$$\hat{SSB} = \sum N_h (\bar{y}_h - \bar{y}_{str})^2$$

$$= 1054(70.9 - 56.7)^2 + 220(8.2 - 56.7)^2 + 1382(38.8 - 56.7)^2 + 442(104.9 - 56.7)^2 \approx 2,201,681$$

Lohr Examples 3.2 and 3.6: Design effect

- The estimated population variance is then

$$\hat{S}^2 = \frac{2201681 + 12926170}{3078 - 1} = 4911.834$$

- The design effect for estimating the population mean `largef92` using this stratified sample is

$$DEff = \frac{3.5577^2}{(1 - 300/3078)4911.834/300} \approx 0.86$$