

Statistical Sampling

Live Session Unit 4

Future Plan

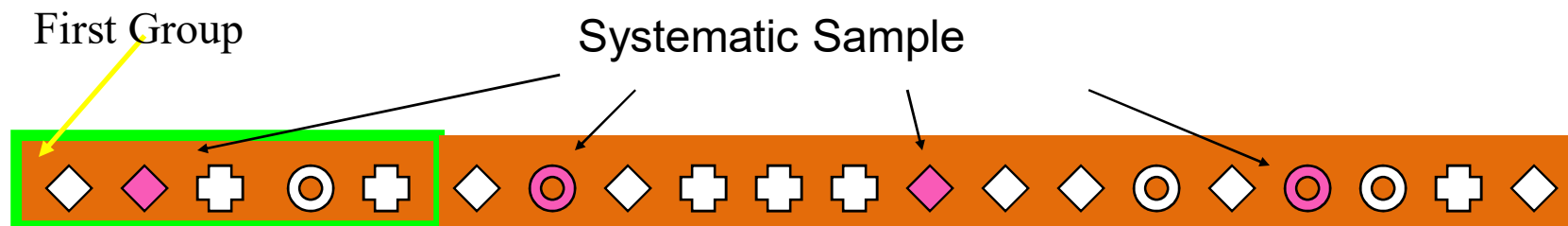
- May 30 : Live session 04- Stratified Design
- Jun 06 : Live session 05- Sample size calculations
- Jun 13 : Live session 06 – Mid term review
- Jun 20 : Mid term

Mid-Term

- Saturday June 17, 9.00a.m CT
- Submit on Tuesday June 20, 11.00p.m. CT
- Text book, Live session notes, Labs, HWs, asynchronous videos (week1-week5).

Systematic Samples

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: $k = N/n$
- Randomly select one individual from the 1st group
- Select every k -th individual thereafter



e.g., $N=300$, $n=55$, $k = 300/55 \approx 5$; Random Starting Position $(1 - 5) = 2$

Exercise!!

A population of 10 balls is made up of 4 white balls and 6 red balls as shown below:

R W W R W R W R R R

A systematic sample of size 2 is selected from the population and the proportion of red balls in the population is estimated from the sample using as the estimator

\hat{p} = the proportion of red balls in the sample

Find the sampling distribution of \hat{p} .

Stratified Design

- Population N units divided into H subpopulations of N_1, N_2, \dots, N_H units.
- These subpopulations are non overlapping.
- These subpopulations are called strata.
- We must know the values of N_1, N_2, \dots, N_H and must have

$$N_1 + N_2 + \dots + N_H = N$$

Notations

- N = Total number of units in the entire population

The following symbols all refer to **stratum h**

- N_h = total number of units
- n_h = number of units in sample
- y_{hi} = value obtained for the *ith* unit
- $W_h = \frac{N_h}{N}$ = stratum weight
- $f_h = \frac{n_h}{N_h}$ = sampling fraction in the stratum

Notations

- The following symbols all refer to **stratum h**

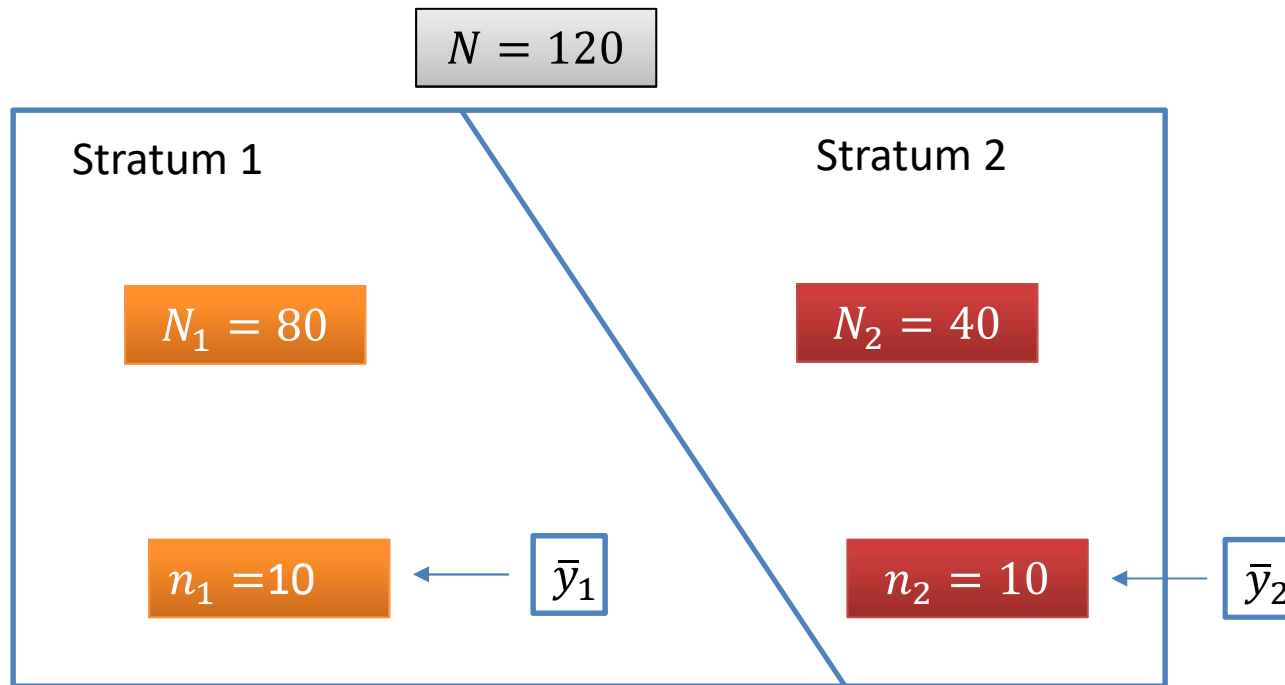
$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h} = \text{True mean}$$

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} = \text{sample mean}$$

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1} = \text{True variance}$$

$$s_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1} = \text{sample variance}$$

Example 01



$$\bar{y}_{str} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \bar{y}_h$$

$$\bar{y}_{st} = \frac{N_1}{N} \bar{y}_1 + \frac{N_2}{N} \bar{y}_2 = \frac{80}{120} \bar{y}_1 + \frac{40}{120} \bar{y}_2$$

Lab 3 discussion

Taxpayers

Taxpayer number	Actual Income (thousands of dollars)
1	60
2	72
3	68
4	94
5	90
6	102
7	116
8	130
9	200

Samples of
size 2 from
pop of size 8

Sample number	2 of 8 taxpayers in sample	mean of actual income of 2 taxpayers \bar{y}_1	Estimate of variance $\hat{\sigma}^2$
1	1,2	66	72
2	1,3	64	32
3	1,4	77	578
4	1,5	75	450
5	1,6	81	882
6	1,7	88	1568
7	1,8	95	2450
8	2,3	70	8
9	2,4	83	242
10	2,5	81	162
11	2,6	87	450
12	2,7	94	968
13	2,8	101	1682
14	3,4	81	338
15	3,5	79	242
16	3,6	85	578
17	3,7	92	1152
18	3,8	99	1922
19	4,5	92	8
20	4,6	98	32
21	4,7	105	242
22	4,8	112	648
23	5,6	96	72
24	5,7	103	338
25	5,8	110	800
26	6,7	109	98
27	6,8	116	392
28	7,8	123	98
Sum		2562	16504
Mean		91.5	589.4286

How does the mean of the sample estimates of mean \bar{y}_1 & var $\hat{\sigma}^2$ compare to the population mean & var?

Mean of 8 Taxpayers actual income = 91.50,

Var of actual income of 8 Taxpayers = 589.43

The mean of the sample estimates of mean \bar{y}_1 and variance $\hat{\sigma}^2$ are equal to the population mean and variance.

How does the mean of the sample estimates of \bar{y}_{st} compare to the population mean?

Mean of 9 Taxpayers actual income = 103.55

The mean of the sample estimates of mean \bar{y}_{st} is equal to the population.

Formula for mean:

$$\bar{y}_{st} = \frac{N_1}{N} \bar{y}_1 + \frac{N_2}{N} \bar{y}_2 = \frac{8}{9} \bar{y}_1 + \frac{1}{9} 200 .$$

Samples of size 3 from Stratified

Sample number	2 of 8 taxpayers in Stratum 1	mean of actual income of 2 taxpayers from Stratum 1, \bar{y}_1	Estimate of population mean of actual income for 9 taxpayers \bar{y}_{st}
1	1,2	66	80.89
2	1,3	64	79.11
3	1,4	77	90.67
4	1,5	75	88.89
5	1,6	81	94.22
6	1,7	88	100.44
7	1,8	95	106.67
8	2,3	70	84.44
9	2,4	83	96.00
10	2,5	81	94.22
11	2,6	87	99.56
12	2,7	94	105.78
13	2,8	101	112.00
14	3,4	81	94.22
15	3,5	79	92.44
16	3,6	85	97.78
17	3,7	92	104.00
18	3,8	99	110.22
19	4,5	92	104.00
20	4,6	98	109.33
21	4,7	105	115.56
22	4,8	112	121.78
23	5,6	96	107.56
24	5,7	103	113.78
25	5,8	110	120.00
26	6,7	109	119.11
27	6,8	116	125.33
28	7,8	123	131.56
Sum			2899.56
Mean			103.56

Stratified Design

- Divide the sampling frame into groups based on stratifying variable(s)
- Select a simple random sample from each stratum
- Put the samples together to estimate

$$\bar{y}_{str} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \bar{y}_h$$

and standard error

$$\hat{\sigma}_{\bar{y}_{str}} = \sqrt{\sum_{h=1}^H \left(1 - \frac{n_h}{N_h} \right) \left(\frac{N_h}{N} \right)^2 \frac{s_h^2}{n_h}}$$

Table 5.2 Payroll Variance

$$h = 1$$

$$\left(1 - \frac{34}{1614}\right) * \left(\frac{1614}{6570}\right)^2 * \frac{183^2}{34} = 58.1906$$

$$h = 2$$

$$\left(1 - \frac{57}{1566}\right) * \left(\frac{1566}{6570}\right)^2 * \frac{316^2}{57} = 95.9068$$

$$h = 3$$

$$\left(1 - \frac{104}{1419}\right) * \left(\frac{1419}{6570}\right)^2 * \frac{641^2}{104} = 170.7891$$

$$h = 4$$

$$\left(1 - \frac{106}{683}\right) * \left(\frac{683}{6570}\right)^2 * \frac{1347^2}{106} = 156.277$$

$$h = 5$$

$$\left(1 - \frac{192}{679}\right) * \left(\frac{679}{6570}\right)^2 * \frac{2463^2}{192} = 242.0447$$

$$h = 6$$

$$\left(1 - \frac{507}{609}\right) * \left(\frac{609}{6570}\right)^2 * \frac{7227^2}{507} = 148.2500$$

$$\sigma_{\bar{x}}^2 = \sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}$$

Total = 871.4582

Stratified Design

estimate total

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

and standard error

$$\hat{\sigma}_{\hat{t}_{str}} = N\hat{\sigma}_{\bar{y}_{str}} = \sqrt{\sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}}$$

Stratified Design

- estimate proportion

$$\hat{p}_{str} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \hat{p}_h$$

and standard error

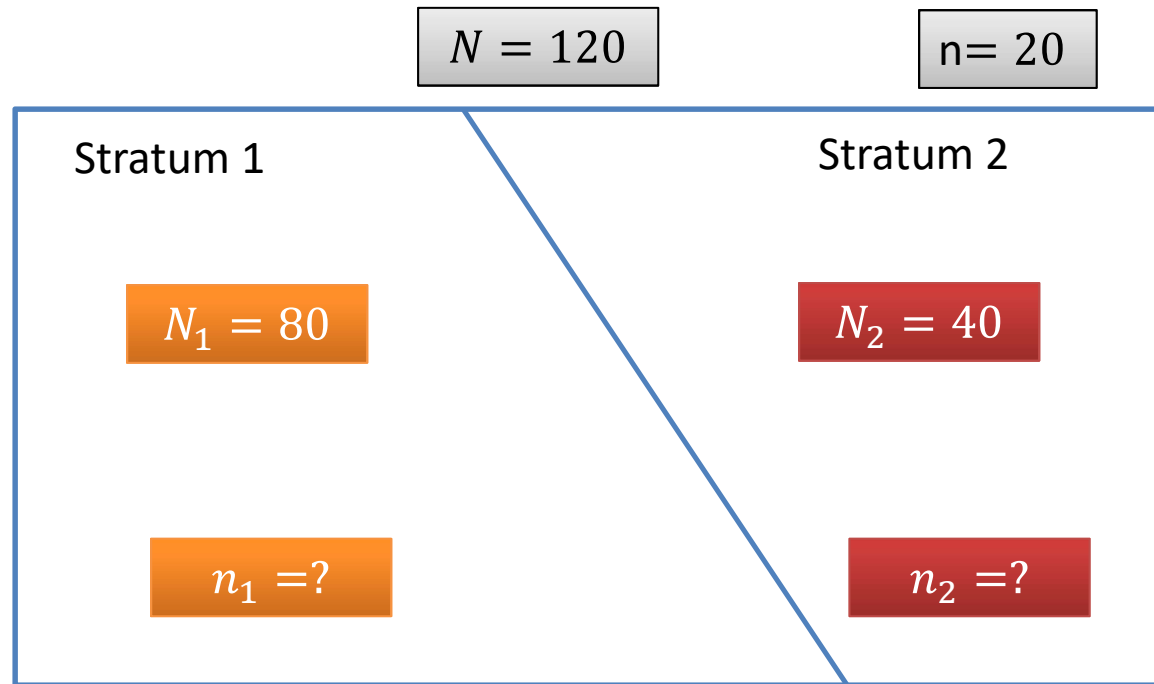
$$\hat{\sigma}_{\hat{p}_{str}} = \sqrt{\sum_{h=1}^H \left(1 - \frac{n_h}{N_h} \right) \left(\frac{N_h}{N} \right)^2 \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}}$$

How do we decide how to allocate our sample among the strata?

- Method 1
 - Proportional allocation
- Proportional allocation takes **sample size proportional to population size** in each stratum;

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

Example 02- Proportional Allocation



$$n_1 = n \left(\frac{N_1}{N} \right) = 20 \left(\frac{80}{120} \right) = 13.33 \approx 13$$

$$n_2 = n \left(\frac{N_2}{N} \right) = 20 \left(\frac{40}{120} \right) = 6.67 \approx 7$$

Example 03

(Source: Sharon Lohr)

- The U.S. government conducts a Census of Agriculture every five years, collecting data on all farms in the 50 states. The census of Agriculture provides data on number of farms, the total acreage devoted to farms, farm size, yield of different crops, and a wide variety of other agricultural measures for each of the counties in United States.
- For this example, we use the four census regions of the United States as strata.

Stratum	N_h	n_h
Northeast	220	
North Central	1054	
South	1382	
West	422	
Total	3078	300

Example 03

(Source: Sharon Lohr)

- The U.S. government conducts a Census of Agriculture every five years, collecting data on all farms in the 50 states. The census of Agriculture provides data on number of farms, the total acreage devoted to farms, farm size, yield of different crops, and a wide variety of other agricultural measures for each of the counties in United States.
- For this example, we use the four census regions of the United States as strata.

Stratum	N_h	n_h
Northeast	220	21
North Central	1054	103
South	1382	135
West	422	41
Total	3078	300

Proportional Allocation - Issues

Texas
Public
Libraries

Stratum	Stratum Size N_h	n_h	Rounded
1	305	41.33	41
2	43	5.83	6
3	15	2.03	2
4	4	0.54	1
5	2	0.27	0
Total	369		50

Problem: cannot estimate variance within strata containing only one unit.

Solutions:

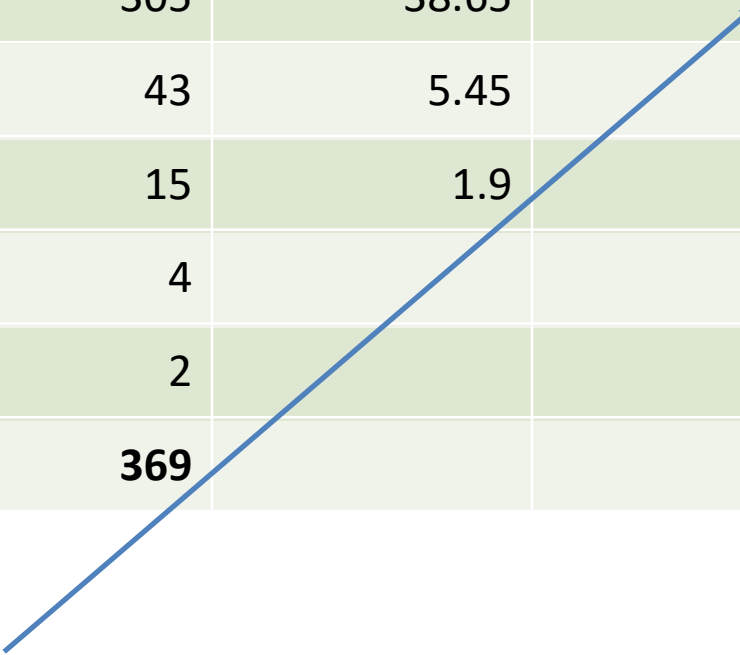
- Collapse strata
- Allocate 2 to each stratum and reallocate proportionately to remaining strata

Proportional Allocation - Issues

Stratum	Stratum Size N_h	n_h	Rounded
1	305		
2	43		
3	15		
4	4		2
5	2		2
Total	369		50

Proportional Allocation - Issues

Stratum	Stratum Size N_h	n_h	Rounded
1	305	38.65	39
2	43	5.45	5
3	15	1.9	2
4	4		2
5	2		2
Total	369		50


$$46 \left(\frac{305}{363} \right)$$

How do we decide how to allocate our sample among the strata?

- Method 2
 - Neyman allocation
- Neyman allocation takes sample size proportional to
stratum size * standard deviation in the stratum:

$$n_h = n \frac{S_h N_h}{\sum_{h=1}^H N_h S_h}$$

- Requires knowledge of standard deviations
- Can use a proxy variable

How do we decide how to allocate our sample among the strata?

Neyman allocation

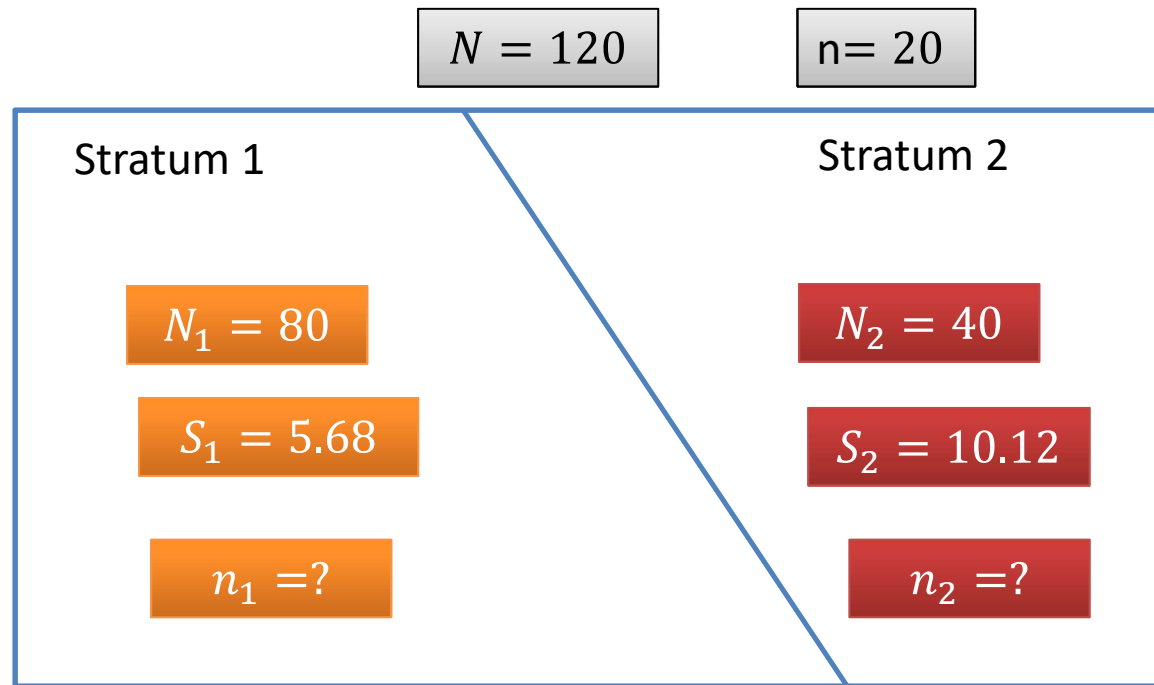
$$n_h = n \frac{S_h N_h}{\sum_{h=1}^H N_h S_h} \quad \text{or} \quad n_h = n \frac{\sigma_h N_h}{\sum_{h=1}^H N_h \sigma_h}$$

In your text book equation (5.2)

$$n_h = n \frac{\sigma_h \pi_h}{\sum_{h=1}^H \pi_h \sigma_h} \quad \text{where } \pi_h = \frac{N_h}{N}$$

$$n_h = n \frac{\sigma_h \frac{N_h}{N}}{\sum_{h=1}^H \frac{N_h}{N} \sigma_h} = n \frac{\sigma_h N_h}{\sum_{h=1}^H N_h \sigma_h}$$

Example 04- Neyman Allocation



$$n_1 = n \frac{S_1 N_1}{\sum_{h=1}^2 N_h S_h} = 20 \left(\frac{5.68 * 80}{5.68 * 80 + 10.12 * 40} \right) = 10.57 \approx 11$$

$$n_2 = n \frac{S_2 N_2}{\sum_{h=1}^2 N_h S_h} = 20 \left(\frac{10.12 * 40}{5.68 * 80 + 10.12 * 40} \right) = 9.42 \approx 9$$

Neyman Allocation - Issues

Stratum	Stratum Size N_h	S_h	$N_h * S_h$	n_h	Rounded
1	305	29701.4	9058627.8	21.21	21
2	43	83225.7	3578703.1	8.38	8
3	15	141059.8	2115897.5	4.95	5
4	4	835320.1	3341280.4	7.82	8
5	2	1628484.6	3256969.2	7.63	8
Total	369				50

Problem: More sample than we have available!

Solution: Take all units from these strata and reallocate to remaining strata.

Neyman Allocation - Issues

Stratum	Stratum Size N_h	S_h	$N_h * S_h$	n_h	Rounded
1	305	29701.4	9058627.8		
2	43	83225.7	3578703.1		
3	15	141059.8	2115897.5		
4	4	835320.1	3341280.4	4	4
5	2	1628484.6	3256969.2	2	2
Total	369				50

Neyman Allocation - Issues

Stratum	Stratum Size N_h	S_h	$N_h * S_h$	n_h	Rounded
1	305	29701.4	9058627.8	27.02	27
2	43	83225.7	3578703.1	10.67	11
3	15	141059.8	2115897.5	6.31	6
4	4	835320.1	3341280.4	4	4
5	2	1628484.6	3256969.2	2	2
Total	369				50

Allocation Comparison

Stratum	Stratum Size N_h	Proportional allocation n_h	Neyman Allocation n_h
1	305	39	27
2	43	5	11
3	15	2	6
4	4	2	4
5	2	2	2
Total	369	50	50

What's so great about Neyman allocation?

- It gives you the smallest possible variance for an estimator of mean or total or proportion of ANY stratified sample with the same n ; in other words, it is the best you can do with stratification
- If the variances are the SAME in every stratum, then Neyman allocation is the same as proportional allocation

Estimation

Goal: How can we use sample data to estimate values of population parameters?

Point estimate: A single statistic value that is the “best guess” for the parameter value

Interval estimate: An interval of numbers around the point estimate, that has a fixed “confidence level” of containing the parameter value. Called a ***confidence interval***.

(Based on sampling distribution and the point estimate)

Confidence Interval

- A **confidence interval** (CI) is an interval of numbers believed to contain the parameter value.
- The probability the method produces an interval that contains the parameter is called the **confidence level**. Most studies use a confidence level such as 0.95 or 0.99.
- Most CIs have the form

point estimate \pm margin of error

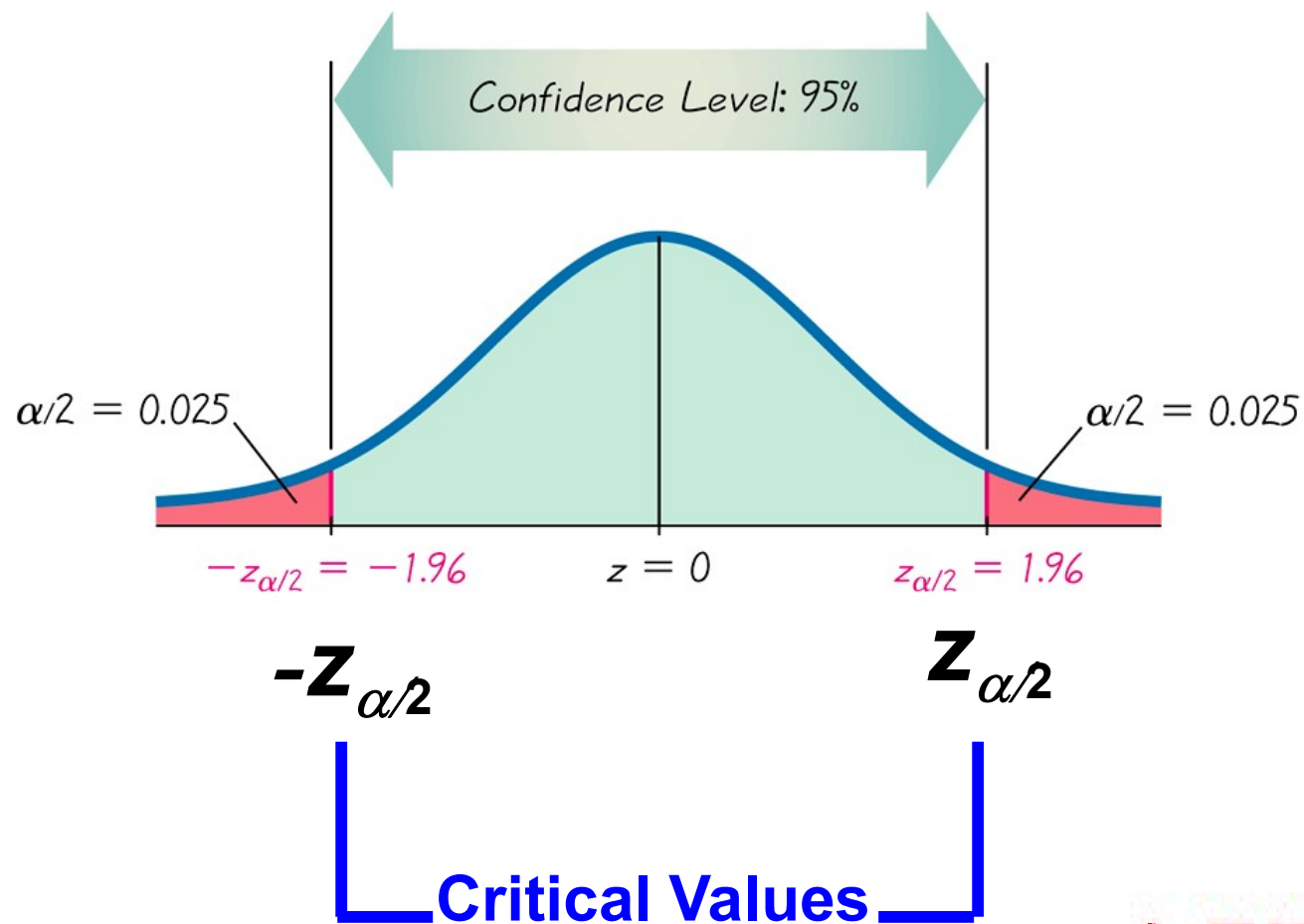
Confidence Interval

point estimate \pm margin of error

$$\bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

$$\text{Confidence Interval} = \left(\bar{x} - Z^* \frac{\sigma}{\sqrt{n}} , \bar{x} + Z^* \frac{\sigma}{\sqrt{n}} \right)$$

Finding $Z_{\alpha/2}$ for a 95% Confidence Level



$$\alpha = 5\%$$

$$\alpha/2 = 2.5\% = .025$$

```
> qnorm(c(0.025, 0.975))  
[1] -1.959964  1.959964
```

Confidence intervals for random samples from FINITE populations

- $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$
- Since we don't know σ , we estimate it by s
- $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} \sqrt{(1 - n/N)}$ is estimated by

$$\hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}} \sqrt{\underbrace{(1 - n/N)}_{\text{fpc}}}$$

- 95% confidence interval for μ is

$$\bar{y} \pm 1.96 * \hat{\sigma}_{\bar{y}}$$

- 95% confidence interval for total is

$$N\bar{y} \pm 1.96 * N\hat{\sigma}_{\bar{y}}$$

Library Example

$$\bar{y} \pm 1.96 * \hat{\sigma}_{\bar{y}}$$

$$\begin{aligned}\hat{\sigma}_{\bar{y}} &= \frac{s}{\sqrt{n}} \sqrt{(1 - n/N)} \\ &= \frac{100,543}{\sqrt{50}} \sqrt{(1 - 50/369)} \\ &= 13220\end{aligned}$$

$$\Rightarrow 24,542 \pm 1.96 * 13220$$

$$\Rightarrow (-1349, 50\ 453)$$

$$\Rightarrow (0, 50\ 453)$$

inq	
Mean	24542.2
Standard Error	14218.88166
Median	800
Mode	0
Standard Deviation	100542.7
Sample Variance	10108829784
Range	682201
Minimum	0
Maximum	682201
Sum	1227109
Count	50

Library Example

$\hat{t}_y \pm 1.96 * \hat{\sigma}_{\hat{t}_y}$, where $\hat{t}_y = N\bar{y} = 9,056,064$ and

$$\begin{aligned}\hat{\sigma}_{\hat{t}_y} &= \frac{Ns}{\sqrt{n}} \sqrt{(1 - n/N)} \\ &= \frac{(369)100,543}{\sqrt{50}} \sqrt{(1 - 50/369)} \\ &= 4,878,362\end{aligned}$$

inq	
Mean	24542.2
Standard Error	14218.88166
Median	800
Mode	0
Standard Deviation	100542.7
Sample Variance	10108829784
Range	682201
Minimum	0
Maximum	682201
Sum	1227109
Count	50

$\Rightarrow 9.056K \pm 1.96 * 4.878K$

$\Rightarrow (0, 18.617K)$

SRS with PROC SURVEYSELECT

To select a SRS using PROC SURVEYSELECT

```
proc surveyselect data = Library out = srssample  
  sampsize = 50 seed = 91114 stats;  
  title "Simple random sample";  
run;
```

```
proc print data = srssample;  
run;
```

Simple random sample


The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
------------------	------------------------

Input Data Set	LIBRARY
Random Number Seed	91114
Sample Size	50
Selection Probability	0.135501
Sampling Weight	7.38
Output Data Set	SRSSAMPLE

Simple random sample

Obs	ID	circ	inq	stratum	SelectionProb	SamplingWeight
1	9	0	0	1	0.13550	7.38
2	14	854	0	1	0.13550	7.38
3	16	1100	75	1	0.13550	7.38
4	21	1619	0	1	0.13550	7.38
5	39	3624	167	1	0.13550	7.38


$$\frac{n}{N} = \frac{50}{369}$$

Allocation Comparison

Stratum	Stratum Size N_h	Proportional allocation n_h	Neyman Allocation n_h
1	305	39	27
2	43	5	11
3	15	2	6
4	4	2	4
5	2	2	2
Total	369	50	50

Stratified sampling with PROC SURVEYSELECT

```
/*Creating strata based on auxiliary variable on the frame*/  
data librarynew;  
set library;  
if circ le 121190 then stratum = 1;  
else if circ le 416226 then stratum = 2;  
else if circ le 853171 then stratum = 3;  
else if circ le 3136876 then stratum = 4;  
else stratum = 5;  
run;
```

Stratified sampling with PROC SURVEYSELECT

```
proc surveyselect data=librarynew method = srs out = str1sample  
  sampsize = (39,5,2,2,2)  
  seed=91115;  
  strata  stratum;  
  title "Proportional allocation";  
run;
```

```
proc surveyselect data=librarynew method = srs out = str2sample  
  sampsize = (27,11,6,4,2)  
  seed=91116;  
  strata  stratum;  
  title "Neyman allocation";  
run;
```

Proportional allocation

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Strata Variable	stratum

Input Data Set	LIBRARYNEW
Random Number Seed	91115
Number of Strata	5
Total Sample Size	50
Output Data Set	STR1SAMPLE

Obs	stratum	ID	circ	inq	SelectionProb	SamplingWeight
1	1	3	0	47	0.12787	7.82051
2	1	5	0	150	0.12787	7.82051
3	1	10	0	0	0.12787	7.82051
4	1	21	1619	0	0.12787	7.82051
5	1	22	2140	15	0.12787	7.82051
6	1	40	3874	150	0.12787	7.82051

45	3	349	442884	99539	0.13333	7.50000
46	3	360	816240	257054	0.13333	7.50000
47	4	365	1705078	682201	0.50000	2.00000
48	4	367	3136876	272711	0.50000	2.00000
49	5	368	4081187	2710130	1.00000	1.00000
50	5	369	6384212	3278281	1.00000	1.00000

Selection Method	Simple Random Sampling
Strata Variable	stratum
Input Data Set	UNIVERSE
Random Number Seed	91115
Number of Strata	5
Total Sample Size	50
Output Data Set	STRSAMPLE

Obs	stratum	ID	circ	SelectionProb	SamplingWeight
1	1	3	0	0.12787	7.82051
2	1	5	0	0.12787	7.82051
3	1	10	0	0.12787	7.82051
4	1	21	1619	0.12787	7.82051
5	1	22	2140	0.12787	7.82051
		...			
46	3	360	816240	0.13333	7.50000
47	4	365	1705078	0.50000	2.00000
48	4	367	3136876	0.50000	2.00000
49	5	368	4081187	1.00000	1.00000
50	5	369	6384212	1.00000	1.00000

Where do those weights come from?

- An estimate of the mean number of inquiries per library as

$$\bar{y}_{str} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \bar{y}_h$$

This estimator can be rearranged to have the form of

$$\bar{y}_{str} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} w_{hi} y_{hi}}{\sum_{h=1}^H \sum_{i=1}^{n_h} w_{hi}},$$

where $w_{hi} = \frac{1}{\pi_{hi}} = \frac{1}{n_h/N_h} = \frac{N_h}{n_h}$

Selection Method	Simple Random Sampling
Strata Variable	stratum
Input Data Set	UNIVERSE
Random Number Seed	91115
Number of Strata	5
Total Sample Size	50
Output Data Set	STRSAMPLE

Obs	stratum	ID	circ	SelectionProb	SamplingWeight
1	1	3	0	0.12787	7.82051 =305/39
2	1	5	0	0.12787	7.82051
3	1	10	0	0.12787	7.82051
4	1	21	1619	0.12787	7.82051
5	1	22	2140	0.12787	7.82051
		...			
46	3	360	816240	0.13333	7.50000
47	4	365	1705078	0.50000	2.00000
48	4	367	3136876	0.50000	2.00000 =4/2
49	5	368	4081187	1.00000	1.00000
50	5	369	6384212	1.00000	1.00000 =2/2

=39/305

=2/4

= 2/2

Analyzing SRS with PROC SURVEYMEANS

```
proc surveymeans data = srssample total = 369 ;  
var inq;  
weight SamplingWeight;  
title "Simple random sample";  
run;
```

Simple random sample

The SURVEYMEANS Procedure

Data Summary

Number of Observations	50
Sum of Weights	369

Statistics

Variable	N	Mean	Std Error of Mean	95% CL for Mean	
inq	50	75823	60871	-46500.861	198147.301

Analyzing SRS with PROC SURVEYMEANS

```
proc surveymeans data = srssample total = 369 mean clm sum  
clsum;  
var inq;  
weight SamplingWeight;  
title "Simple random sample";  
run;
```

Simple random sample

The SURVEYMEANS Procedure


Data Summary	
Number of Observations	50
Sum of Weights	369

Statistics								
Variable	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Dev	95% CL for Sum	
inq	75823	60871	-46500.861	198147.301	27978768	22461257	-17158818	73116354.2

Analyzing Stratified sample with PROC SURVEYMEANS

Stratum	Stratum Size N_h
1	305
2	43
3	15
4	4
5	2
Total	369

```
data strsizes;  
input stratum _total_;  
datalines;  
1 305  
2 43  
3 15  
4 4  
5 2  
;  
run;
```



```
proc surveymeans data = str1sample mean clm sum clsum total = strsizes;  
var inq;  
weight SamplingWeight;  
strata stratum;  
title "Proportional allocation";  
run;
```

Results for proportionately allocated sample

Proportional allocation

The SURVEYMEANS Procedure


Data Summary	
Number of Strata	5
Number of Observations	50
Sum of Weights	369

Statistics								
Variable	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Dev	95% CL for Sum	
inq	33763	3589.379522	26534.0586	40992.8215	12458709	1324481	9791067.62	15126351.1

Analyzing Stratified sample with PROC SURVEYMEANS

Stratum	Stratum Size N_h
1	305
2	43
3	15
4	4
5	2
Total	369

```
data strsizes;  
input stratum _total_;  
datalines;  
1 305  
2 43  
3 15  
4 4  
5 2  
;  
run;
```



```
proc surveymeans data=str2sample mean clm sum clsum total = strsizes;  
var inq;  
weight SamplingWeight;  
strata stratum;  
title "Neyman allocation";  
run;
```

Results for Neyman allocated sample

Neyman allocation

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	5
Number of Observations	50
Sum of Weights	369

Statistics								
Variable	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Dev	95% CL for Sum	
inq	25029	765.613863	23487.0769	26571.1278	9235739	282512	8666731.37	9804746.17

Comparison of three designs in library example

Sample design	Estimate	Stderr
SRS	27,978,768	10,961,257
STR, proportionate	12,457,169	1,324,500
STR, Neyman	9,235,739	282,512

True value of total inquiries = 10,704,683

The value of stratification...

- You get “more bang for your buck” with each sample unit than from a SRS; i.e., smaller variance for your estimator
- To be able to take advantage of this, you must have in advance of sampling, information about each sample unit that allows you to divide them into strata that are homogenous, with respect to the main variable of interest.
- To use Neyman allocation, you must have even MORE information, but proxy variables (such as measures of size) often do a very good job.
- No matter how you form strata, if you sample with equal probability in each stratum, your estimator will not have a higher variance than it would have from a SRS of the same size

Watch Video 4.7 again!

A practical issue we have not addressed is....

- How big a sample size do we need?
- Clearly, it will differ depending on the design we plan