Project: Details in Canvas; note 1^{st} milestone

STAT 466: Survey Sampling

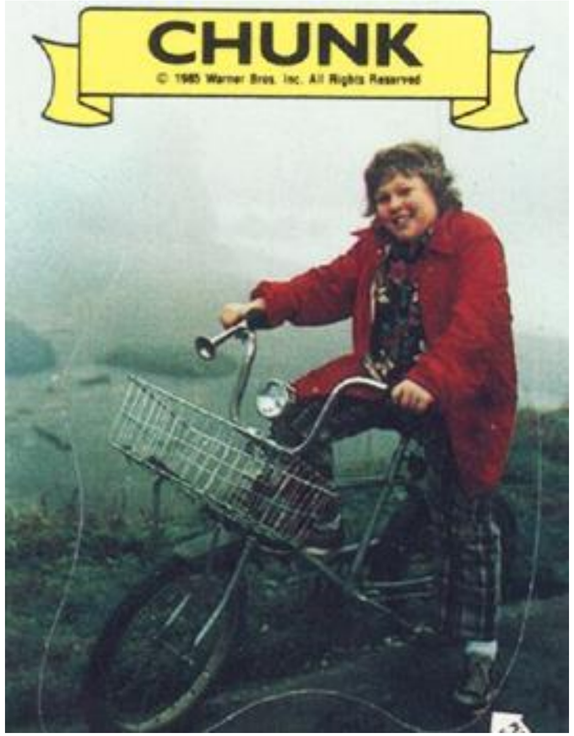
GOAL: Estimate the population mean passenger age and place a bound on the error of estimation.

Ch5 Stratified Random Sample

(sample size, assumptions, estimate)

Stratified Random Sampling

Strata:
Nonoverlapping groups that a population is partitioned into



Stratified random sampling:
Sample selected within each stratum using simple random sampling

Stratified Random Sampling: Defining Strata

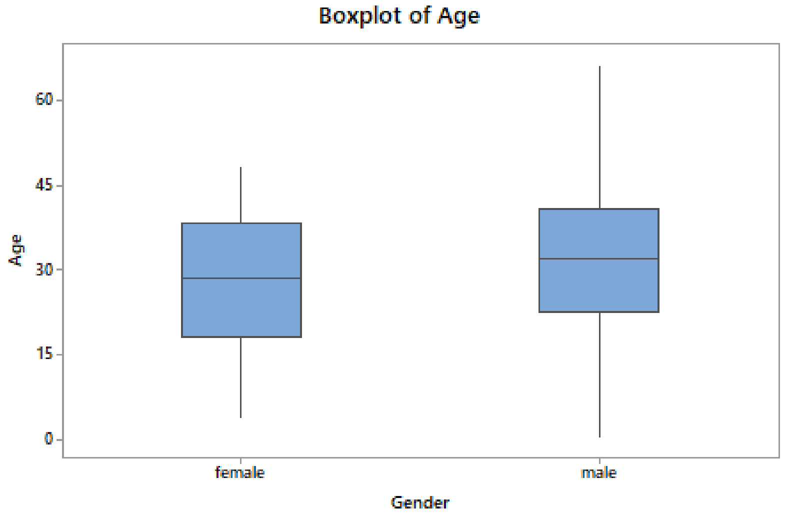
Suppose you consider stratifying by Class:

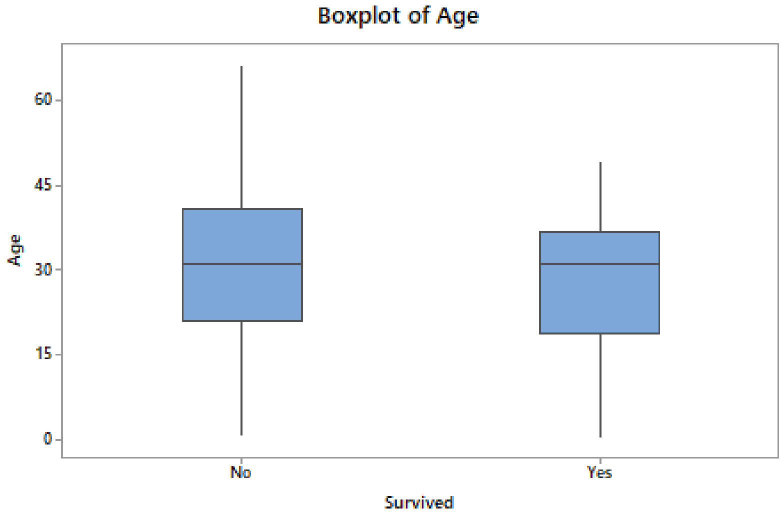
320 1st class passengers

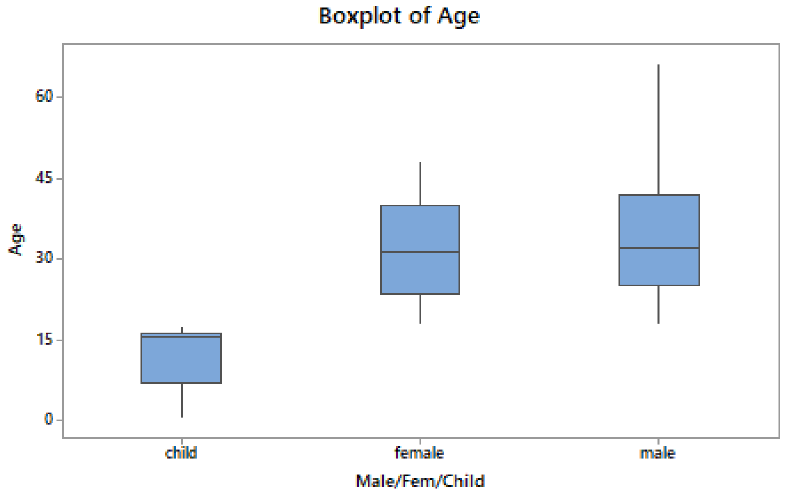
285 2nd class passengers

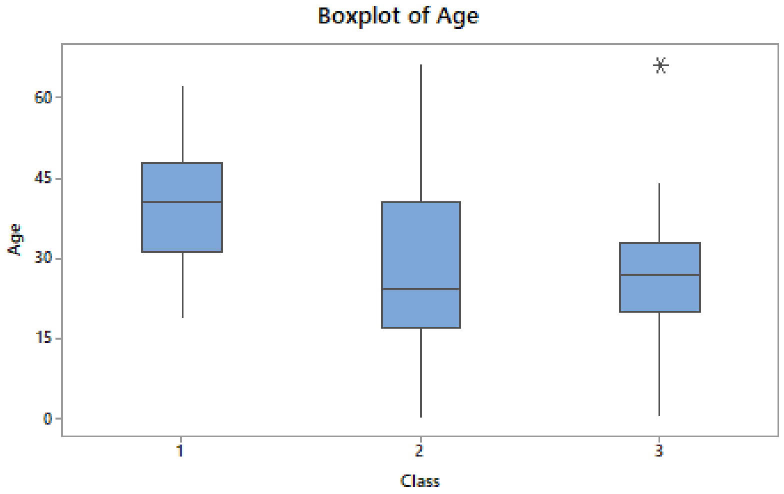
706 3rd class passengers

Stratified Random Sampling: Defining Strata









Worksheet: TitanicSample_130n.mtw

These are boxplots of SRS data (OR for your projects, use the population).

Which variable should we use to define our strata?

Sample size using Neyman allocation

If the cost per observation is the same (or unknown) for all strata, then we can use **Neyman allocation** to calculate n_i

Neyman Allocation ▸ Step 1

For N=1311 passengers, suppose we want a bound of 2 years and know the following strata details:

	age range	*#passengers*
children	*17 years*	*183*
females	*46 years*	*384*
males	*56 years*	*744*

Neyman Allocation ▸ Step 1

Using the details from the previous slide:

$$n = \frac{\left(\sum_{k=1}^{L} N_k \sigma_k\right)^2}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2} \approx 127 \text{ passengers}$$

$D = \frac{B^2}{4}$ *when estimating* μ

$D = \frac{B^2}{4N^2}$ *when estimating* τ

Neyman Allocation ▸ Step 1

OR for N=1311 passengers, suppose we want a bound of 2 years and know the following strata details:

	stdev	*#passengers*
children	*5.9 years*	*183*
females	*11.7 years*	*384*
males	*11.8 years*	*744*

Neyman Allocation ▸ Step 1

Using the details from the previous slide:

$$n = \frac{\left(\sum_{k=1}^{L} N_k \sigma_k\right)^2}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2} \approx 110 \text{ passengers}$$

$D = \frac{B^2}{4}$ *when estimating* μ

$D = \frac{B^2}{4N^2}$ *when estimating* τ

Neyman Allocation ▸ Step 2

$$n_i = n\left(\frac{N_i\sigma_i}{\sum_{k=1}^{L} N_k\sigma_k}\right)$$

To estimate μ with a 2 year bound:

$$n_1 = na_1 = 110(0.0752) = 8.3 \approx 8$$

$$n_2 = na_2 = 110(0.3131) = 34.4 \approx 34 \; or \; 35$$

$$n_3 = na_3 = 110(0.6117) = 67.3 \approx 67$$

Stratified Random Sampling: Mean Estimate with a Bound

$$\bar{y}_{\text{st}} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$$

$$\hat{V}(\bar{y}_{\text{st}}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

Using stratified random sampling:

- Did we get the bound we wanted?
- Did our ~95% confidence interval capture the true mean $\mu = 29.392$?

Ch4 Simple Random Sample

(sample size, assumptions, estimate)

SRS: Mean Estimate with a Bound

Let's then randomly sample n passengers…

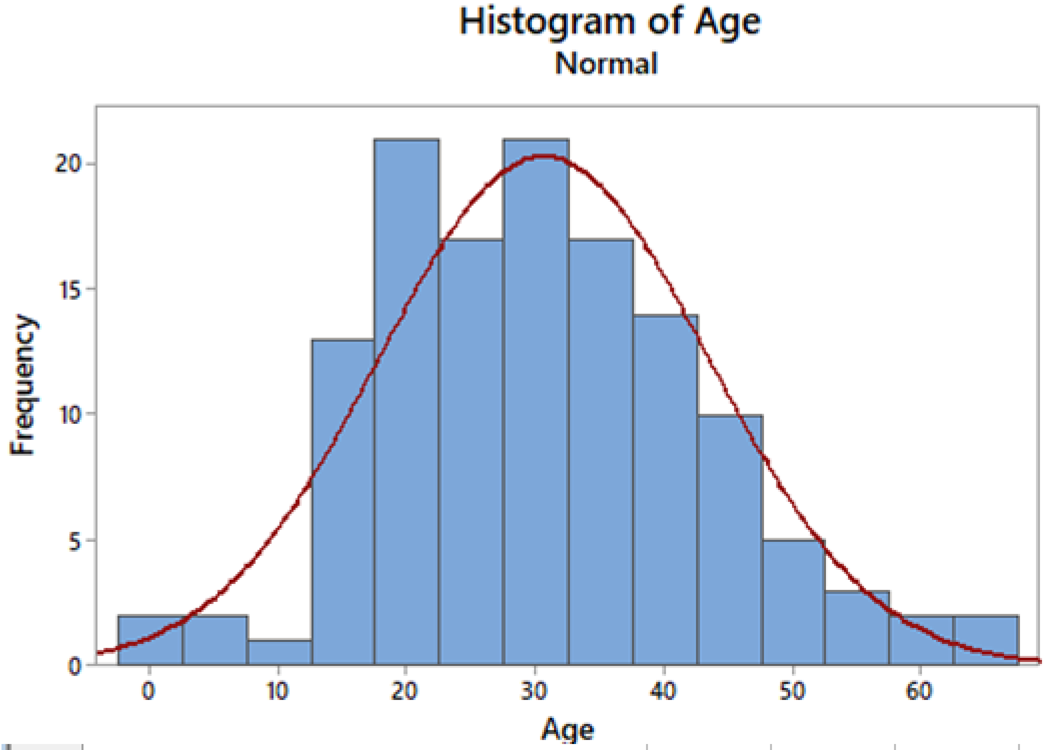
(Minitab: Calc > Random Data > Sample from Columns)

Titanic (1).mtw ***

↓	C1-T	C2
	Name	Age
1	Miss Elisabeth Walton ALLEN	29.0
2	Mr Hudson Joshua Creighton ALLISON	30.0
3	Mrs Bessie Waldo ALLISON	25.0
4	Miss Helen Loraine ALLISON	2.0
5	Master Hudson Trevor ALLISON	0.9
6	Mr Harry ANDERSON	47.0
7	Miss Kornelia Theodosia ANDREWS	62.0
8	Mr Thomas ANDREWS	39.0
9	Mrs Charlotte APPLETON	53.0
10	Mr Ramon ARTAGAVEYTIA	71.0
11	Colonel John Jacob ASTOR	47.0
12	Mrs Madeleine Talmage ASTOR	18.0
13	Miss Léontine Pauline AUBART	24.0
14	Miss Ellen "Nellie" BARBER	26.0
15	Mr Algernon Henry BARKWORTH	47.0
16	Mr John D. BAUMANN	60.0
17	Mrs Hélène BAXTER	50.0
18	Mr Quigg Edmond BAXTER	24.0

SRS: Mean Estimate with a Bound

Given our sample data, are the assumption(s) satisfied?



SRS: Mean Estimate with a Bound

$$\bar{y} = \frac{\sum y_i}{n}$$

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

Using SRS:

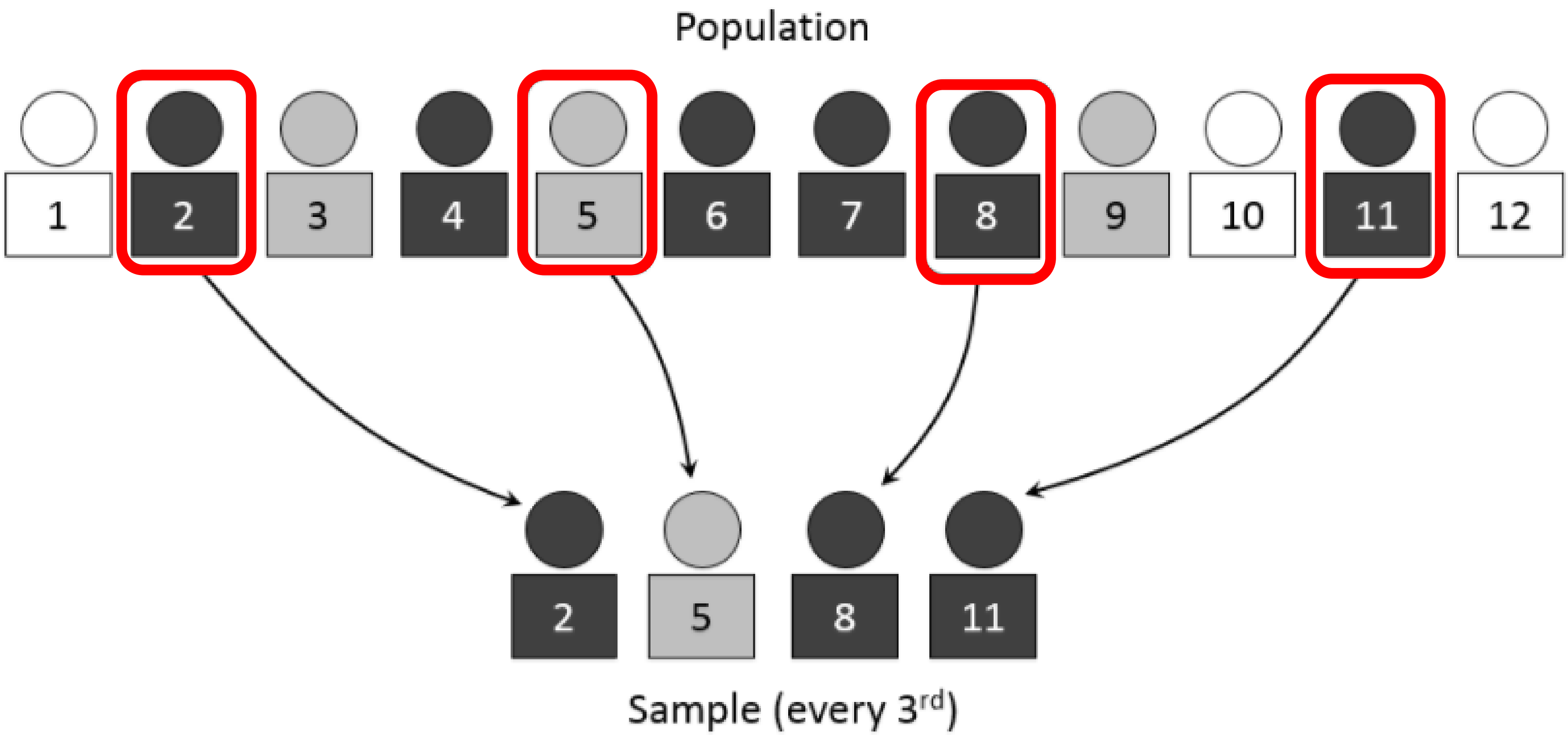
- Did we get the bound we wanted?
- Did our ~95% confidence interval capture the true mean $\mu = 29.392$?

Ch7 Systematic Sample

(sample size, assumptions, estimate)

1-in-*k* Systematic Sampling

E.g. randomly select starting point, then select every 3rd name thereafter



Systematic Sample: Sample Size for a Mean

If n=110, then what is k given a random starting point?

$$\frac{N}{n} = \frac{1311}{110} \leq 11.9$$

Systematic Sample: Mean Estimate with a Bound

Let's then randomly sample every 11th passenger, given a random starting point…

Titanic (1).mtw ***

↓	C1-T	C2	C3	C4-T	C5-T	C6-T	C7
	Name	Age	Class	Gender	Survived	Male/Fem/Child	1:k
1	Miss Elisabeth Walton ALLEN	29.0	1	female	Yes	female	1
2	Mr Hudson Joshua Creighton ALLISON	30.0	1	male	No	male	2
3	Mrs Bessie Waldo ALLISON	25.0	1	female	No	female	3
4	Miss Helen Loraine ALLISON	2.0	1	female	No	child	4
5	Master Hudson Trevor ALLISON	0.9	1	male	Yes	child	5
6	Mr Harry ANDERSON	47.0	1	male	Yes	male	6
7	Miss Kornelia Theodosia ANDREWS	62.0	1	female	Yes	female	7
8	Mr Thomas ANDREWS	39.0	1	male	No	male	8
9	Mrs Charlotte APPLETON	53.0	1	female	Yes	female	9
10	Mr Ramon ARTAGAVEYTIA	71.0	1	male	No	male	10
11	Colonel John Jacob ASTOR	47.0	1	male	No	male	11
12	Mrs Madeleine Talmage ASTOR	18.0	1	female	Yes	female	1
13	Miss Léontine Pauline AUBART	24.0	1	female	Yes	female	2

Systematic Sample: Mean with a Bound

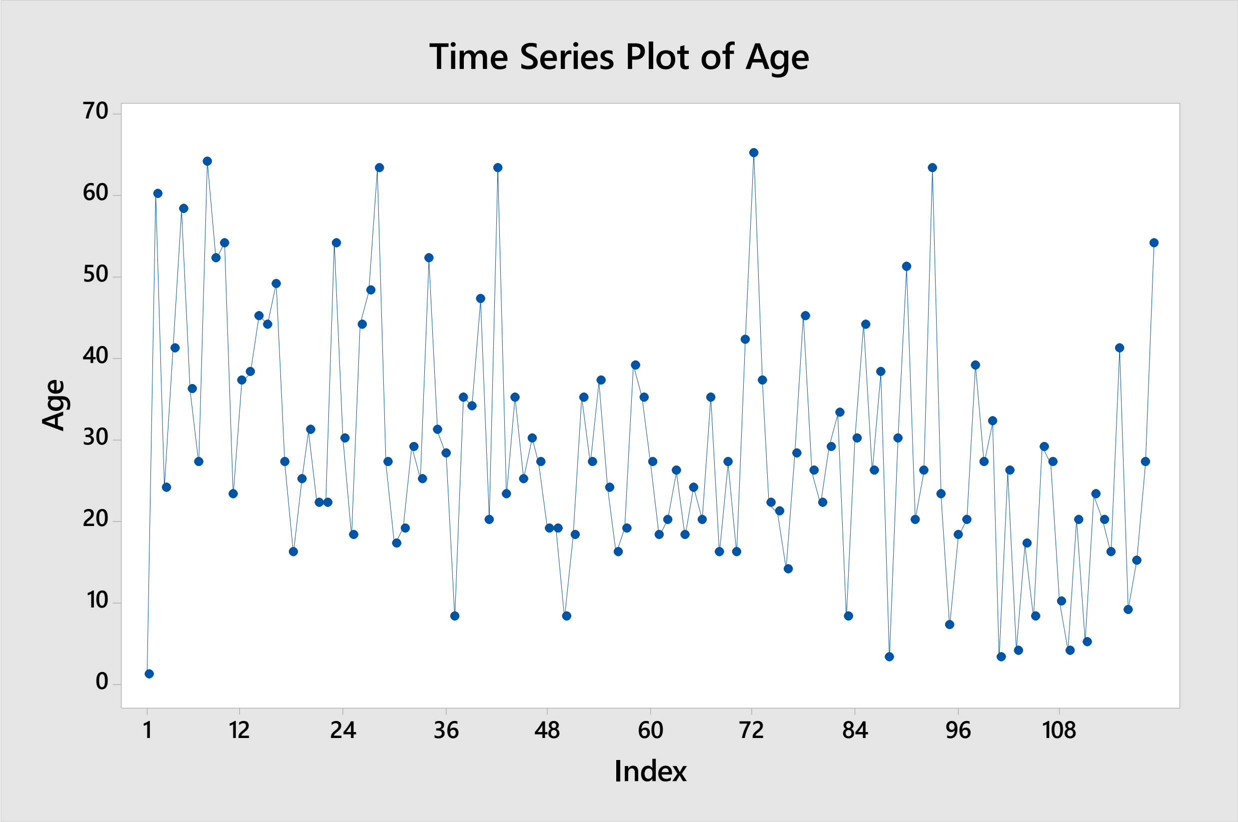
Suppose our random start is the 5th passenger in the population frame:

Systematic sample ***

↓	C1-T	C2	C3	C4-T	C5-T	C6-T	C7
	Name	Age	Class	Gender	Survived	Male/Fem/Child	1:k
1	Master Hudson Trevor ALLISON	0.9	1	male	Yes	child	5
2	Mr John D. BAUMANN	60.0	1	male	No	male	5
3	Mr Jakob BIRNBAUM	24.0	1	male	No	male	5
4	Mr John Bertram BRADY	41.0	1	male	No	male	5
5	Mrs Charlotte Wardle CARDEZA	58.0	1	female	Yes	female	5
6	Mr Tyrell William CAVENDISH	36.0	1	male	No	male	5
7	Mr Walter Miller CLARK	27.0	1	male	No	male	5
8	Mrs Catherine Elizabeth CROSBY	64.0	1	female	Yes	female	5
9	Dr Washington DODGE	52.0	1	male	Yes	male	5
10	Miss Elizabeth Mussey EUSTIS	54.0	1	female	Yes	female	5
11	Miss Mabel Helen FORTUNE	23.0	1	female	Yes	female	5
12	Mr Jacques Heath FUTRELLE	37.0	1	male	No	male	5

Systematic Sample: Mean Estimate with a Bound

Are the assumption(s) satisfied?



Systematic Sample: Mean with a Bound

$$\bar{y}_{sy} = \frac{\sum y_i}{n}$$

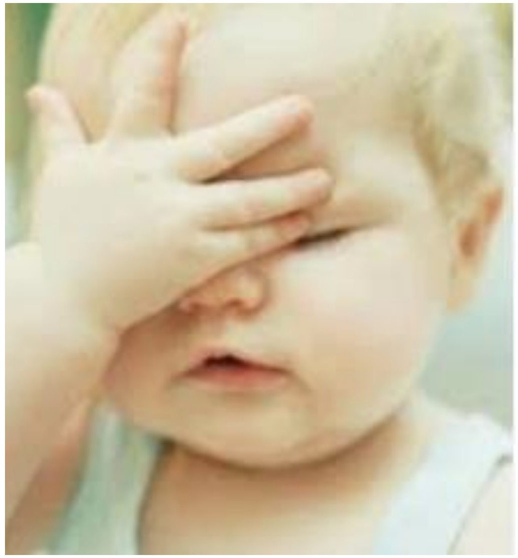
$$\hat{V}(\bar{y}_{sy}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

Using a systematic sample:

- Did we get the bound we wanted?
- Did our ~95% confidence interval capture the true mean $\mu = 29.392$?

Systematic Sample: Mean Estimate with a Bound

Thinking back to the assumptions, what is we think there is a trend?



Successive Differences

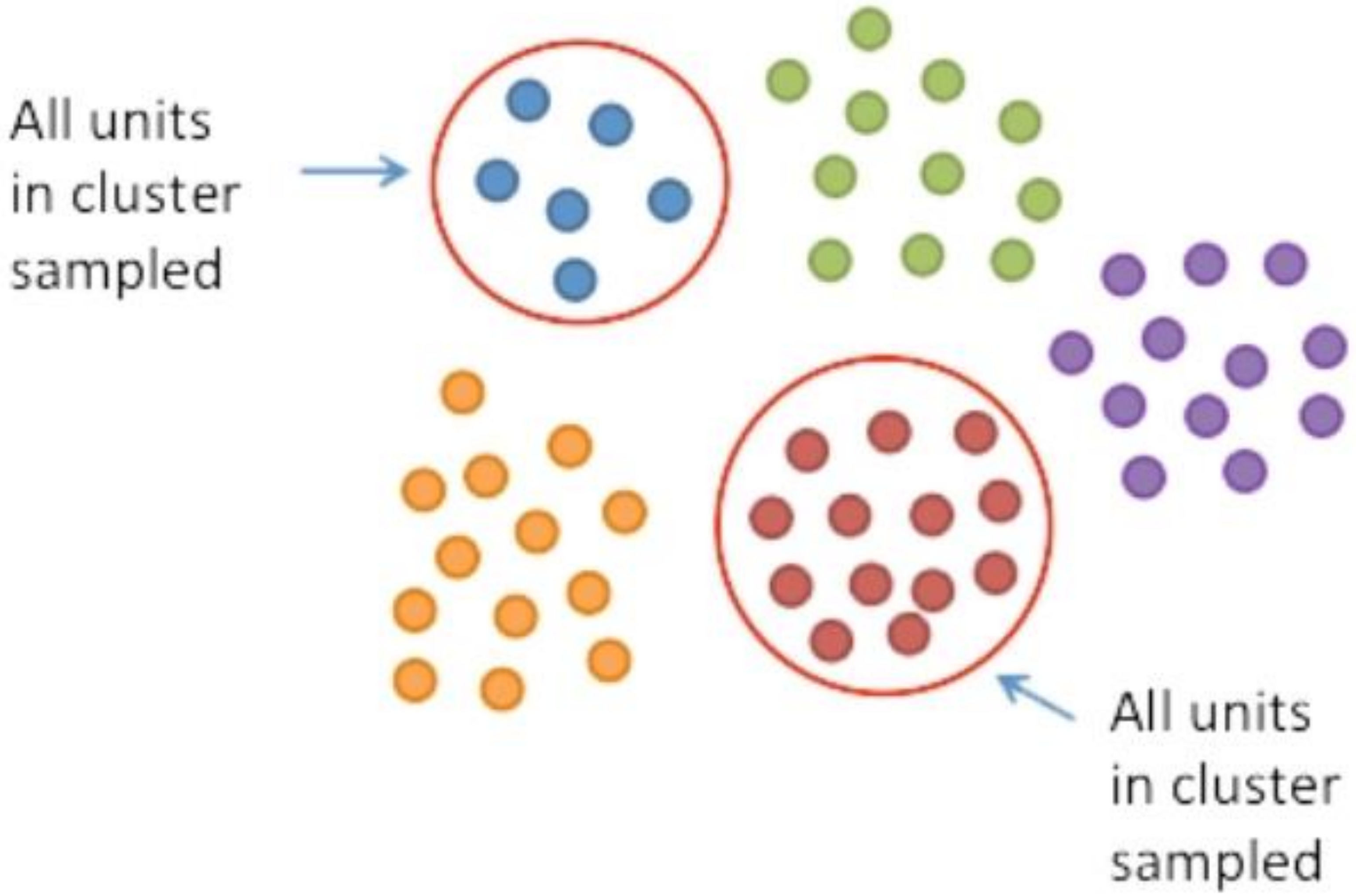
$$d_i = y_{i+1} - y_i, \qquad i = 1, \ldots, (n-1)$$

$$\hat{V}_d(\bar{y}_{sy}) = (1 - \frac{n}{N}) \frac{1}{2n(n-1)} \sum_{i=1}^{n-1} d_i^2$$

Ch8 Cluster Sample

(sample size, assumptions, estimate)

Cluster Sampling



Cluster Sampling: Defining Clusters

- Which variable should we use to define our clusters?

- You could even consider **combining variables** to form clusters (e.g. Class & Gender, which would yield 6 clusters from which to randomly select)

Cluster Sampling: Defining Clusters

Cluster Sampling: Mean with a Bound

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

Estimated variance of $\bar{y}$:

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_r^2}{n\bar{M}^2}$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}m_i)^2}{n - 1}$$

Recall: Estimate $\bar{M} \approx \bar{m}$ if M is unknown

Ch9 Two-Stage Cluster Sample

Two-Stage Cluster Sampling

First ➤ Obtain a frame listing all clusters in the population > select a SRS of clusters

Then ➤ Obtain frames listing all elements for each sampled cluster > select a SRS of elements from each of these frames

Two-Stage Cluster μ :

$$\hat{\mu} = \left(\frac{N}{M}\right)\frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n} = \frac{1}{\bar{M}}\frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n}$$

N = number of clusters in the population
n = number of clusters selected in a SRS
M_i = number of elements in cluster i
M = number of elements in the population ($\sum M_i$)
$\bar{M} = M/N$ = average cluster size for the population
$\bar{y}_i$ = sample mean for i^{th} cluster

Two-Stage Cluster Var. of $\widehat{\mu}$:

$$\hat{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n\overline{M}^2}\right)s_b^2 + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2\left(1 - \frac{m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right)$$

where

$$s_b^2 = \frac{\sum_{i=1}^{n}(M_i\overline{y}_i - \overline{M}\hat{\mu})^2}{n - 1}$$

and

$$s_i^2 = \frac{\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_i)^2}{m_i - 1} \qquad i = 1, 2, \ldots, n$$

$m_i =$ number of elements selected in a SRS from cluster i