STA304/1003 H1 F - Summer 2014: Surveys, Sampling, and Observational Data

Lecture 3 - Part II: Simple Random Sampling

Ramya Thinniyam

May 22, 2014

Simple Random Samples

Recall - Simple Random Sampling without Replacement (SRS) :

- Population: $\mathcal{U} = \{1, 2, \ldots, N-1, N\}$, Sample: $\mathcal{S}$ of size n
- $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ possible samples
- Each sample has probability of $\frac{1}{\binom{N}{n}}$ of being selected
- $\Rightarrow$ $\pi_i = P(\textit{ith unit is selected}) = \frac{n}{N}$
- Each sampling unit is in $\binom{N-1}{n-1}$ different samples
- Successive draws are NOT independent - probabilities change after each draw

To obtain a SRS:

1. List of all units in the population (sampling frame). The sampling unit=observational unit
2. Assign each unit a number between 1 and N
3. Select sample so that each possible sample of size n has the same chance of selection. Use:
   - Random Number Generator
   - 'R' or any software

Example: Using R to Generate SRSs

$N = 25, n = 10$:

```
> myvariable
 [1]  3.60 8.31 5.13 0.66 3.93 4.66 9.04 5.86 0.96 8.60 7.93 4.25
[13] 3.12 2.98 1.25 2.42 3.32 6.53 0.63 0.56 2.04 4.95 4.61 3.51 5.37
```

Sample units:

```
> sample(1:25,10,replace=F)
 [1] 23 13 18 17 24  7 20 12  2  8

> sample(1:25,10,replace=T)
 [1] 13 24 17 13 13 10 10  3  3 21
```

F: without replacement

T: with replacement

repeated

don't need T

Sample variable:

```
> sample(myvariable,10,replace=F)
[1] 2.98 0.63 3.51 4.61 5.86 3.60 5.13 0.66 4.66 1.25

> sample(myvariable,10,replace=T)
[1] 8.60 6.53 2.42 0.56 3.32 8.60 2.42 2.42 0.96 3.12
```

OR

```
> units <- sample(1:25,10,replace=F)
> units
 [1] 21 24  1 13 11  7 12  9 23 10

> myvariable[units]
2.04 3.51 3.60 3.12 7.93 9.04 4.25 0.96 4.61 8.60
```

We don't want T

```
> units <- sample(1:25,10,replace=T)
> units
 [1] 19  5  3  3 20 11 21  4 25 17

> myvariable[units]
0.63 3.93 5.13 5.13 0.56 7.93 2.04 0.66 5.37 3.32
```

Population Parameters and Estimates

- Population Total: $t = \sum_{i=1}^{N} y_i$

  Estimate: $\hat{t} = N\bar{y}_S$ → depends on sample

- Population Mean: $\bar{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i$

  Estimate: $\bar{y}_S = \frac{1}{n}\sum_{i\in S} y_i = \bar{y}$ estimate of pop mean is sample mean

- Variance of Population Values: $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2$

  Estimate with sample variance: $s^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \bar{y})^2$

- Proportion:

  $p = \frac{\#\ of\ units\ in\ population\ with\ desired\ characteristic}{N} = \bar{y}_U$;

  if letting $y_i = I(unit\ i\ has\ characteristic)$

  Estimate: $\hat{p} = \frac{\#\ of\ units\ in\ sample\ with\ desired\ characteristic}{n} = \bar{y}$

  (typo here)

Properties of $\hat{t}$ and $\bar{y}$

- Expected Values:

$$E(\hat{t}) = t \quad \text{and} \quad E(\bar{y}) = \bar{y}_U$$

Both unbiased estimators

only difference : N^2

- Variances:

$$V(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \quad \text{and} \quad V(\bar{y}) = \frac{S^2}{n} \left(1 - \frac{n}{N}\right)$$

The term $1 - \frac{n}{N}$ is called the finite population correction (fpc): small populations give greater sampling fraction $\frac{n}{N}$ so we have more information about the population and so smaller variance. Recall that $V(\bar{y})$ and $V(\hat{t})$ measure variability among estimates of $\bar{y}_U$ and t from different samples.

We need to estimate S^2 using the unbiased estimator s^2 which yields: $\hat{V}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$ and $\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$

PROOFS

Remember what is random and what is NOT . . .

SRS property of $\hat{t}$ & $\bar{y}$:

Proofs: y_i are fixed, but unknown, the RVs are which units get selected in sample.

$Z_i = I$ (unit i selected)

Already showed $Z_i \overset{\text{i.d. not iid (b/c not indep. but identically)}}{\sim} \text{Bern}(n/N)$, $E(Z_i) = \frac{n}{N}$, $\text{Var}(Z_i) = \frac{n}{N}(1-\frac{n}{N})$ (NOT independent)

fpc factor

$$E(\bar{y}) = E\left[\frac{1}{n}\sum_{i\in S} y_i\right] = \frac{1}{n} E\left(\sum_{i=1}^{N} Z_i y_i\right) = \frac{1}{n}\sum_{i=1}^{N} y_i E(Z_i) = \frac{1}{n}\sum_{i=1}^{N} y_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U$$

(fixed => constant; Z_i variable)

Trick

$$\text{Var}(\bar{y}) = V\left(\frac{1}{n}\sum_{i\in S} y_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{N} Z_i y_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i\neq j} y_i y_j \text{Cov}(y_i, y_j)\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \frac{n}{N}(1-\frac{n}{N}) - \sum_{i\neq j} y_i y_j \frac{1}{N-1}(1-\frac{n}{N})\frac{n}{N}\right]$$

$$= \frac{1}{n^2}\frac{n}{N}(1-\frac{n}{N})\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i\neq j} y_i y_j\right]$$

$$= \frac{1}{n}(1-\frac{n}{N})\frac{1}{N(N-1)}\left[(N-1)\sum_{i=1}^{N} y_i^2 - (\Sigma y_i)^2 + \Sigma y_i^2\right] \text{ by } \circledast$$

$$= \frac{1}{n}(1-\frac{n}{N})\frac{1}{N(N-1)}\left[N\Sigma y_i^2 - (\Sigma y_i)^2\right] = \frac{1}{n}(1-\frac{n}{N})\frac{N}{N(N-1)}\left[\sum_{i=1}^{N} y_i^2 - \frac{(\Sigma y_i)^2}{N}\right] = \frac{1}{n}(1-\frac{n}{N})S^2$$

note: $(\Sigma y_i)^2 = \Sigma y_i^2 + \sum_{i\neq j} y_i y_j$

$\circledast$ polynomial expansion

Since $\hat{t} = N\bar{y}$

$$E(\hat{t}) = N E(\bar{y}) = N\bar{y}_U = t$$

$$V(\hat{t}) = \text{Var}(N\bar{y}) = N^2 \text{Var}(\bar{y}) = N^2(1-\frac{n}{N})\frac{S^2}{N}$$

Example: $N = 100, n = 10$ using R

```
> mean(mydata)
[1] 9.874
> var(mydata)
[1] 10.03023
```

'mydata' is actually a vector of y values in population $= (y_1, \cdots, y_{100})$

$\bar{y}_U = 9.874$ $S^2 = 10.03023$

$=$ pop. mean $=$ pop. var.

a) What is the population mean and variance?

b) Find the expected value and variance of $\bar{y}$ for the following samples:

(i) Sample 1:

```
> sample1<-sample(mydata,10,replace=F)
> mean(sample1)
[1] 10.14
> var(sample1)
[1] 9.296
```

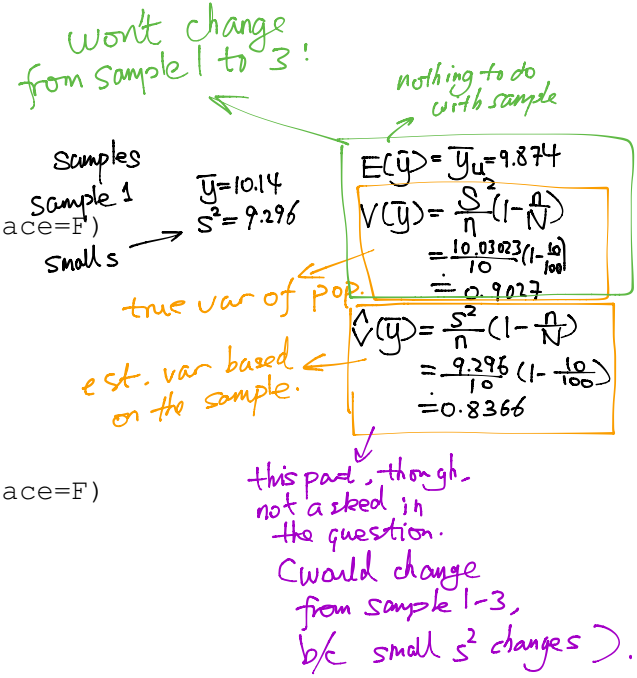
Example (con'd)...

(ii) Sample 2:

```
> sample2<-sample(mydata,10,replace=F)
> mean(sample2)
[1] 7.38
> var(sample2)
[1] 8.237333
```

(iii) Sample 3:

```
> sample3<-sample(mydata,10,replace=F)
> mean(sample3)
[1] 11.08
> var(sample3)
[1] 12.16178
```



Other Measures

Standard Error (SE): square root of the estimated variance of the estimator.

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

Coefficient of Variation (CV): measure of relative variability.

$$CV(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})};$$

for $E(\hat{\theta}) \neq 0$.

- Population standard deviation is often related to mean
- CV doesn't depend on unit of measurement (advantage)

$$CV(\bar{y}) = \sqrt{1 - \frac{n}{N}}\frac{S}{\sqrt{n}\bar{y}_U};$$

for $\bar{y}_U \neq 0$.

Estimated Coefficient of Variation: standard error of estimator divided by mean (standard error expressed as a percentage of the mean).

$$\widehat{CV}(\bar{y}) = \frac{SE(\bar{y})}{\bar{y}} = \sqrt{1 - \frac{n}{N}}\frac{s}{\sqrt{n}\bar{y}} \text{ ; for } \bar{y}_U \neq 0$$

Note: $CV(\hat{t}) = CV(\bar{y})$. b/c $\hat{t} = N\bar{y}$, variability is fixed.

Estimating a Proportion

$y_i = I(\text{unit } i \text{ has characteristic})$ pre-condition !

- $p = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U$
- $\hat{p} = \bar{y}$
- $\hat{p}$ is unbiased for p
- $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - p)^2 = \frac{N}{N-1}p(1-p)$
- $V(\hat{p}) = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}$
- $s^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \hat{p})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$
- $\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$

PROOFS

basically same as $\bar{y}$, $\hat{t}$ proof above.

Recall that $\hat{p}$ is a sample mean and y_i are binary data . . .

Proportion Proofs:

- $p = \sum_{i=1}^{N} y_i / N = \bar{y}_U$

  $y_i = I(i \text{ has characteristic})$
- $\hat{p} = \bar{y}$ is unbiased for $\bar{y}_U = p$. done
- capital $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - p)^2 = \frac{1}{N-1} (\sum^N y_i^2 - 2p \sum^N y_i + Np^2) = \frac{1}{N-1} (\sum^N y_i - 2p \sum y_i + Np^2)$

  recall $y_i = y_i^2$ as "indicator"

  $= \frac{1}{N-1} [(1-2p) \sum y_i + Np^2]$

  $= \frac{1}{N-1} [(1-2p) \cdot pN + Np^2]$

  $= \frac{N}{N-1} p(1-p)$
- $V(\hat{p}) = V(\bar{y}) = \frac{S^2}{n}(1 - \frac{n}{N}) = \frac{N}{N-1} \frac{p(1-p)}{n} \cdot \frac{N-n}{N} = \frac{N-n}{N-1} \frac{p(1-p)}{n}$

HW: do almost same to small s^2

& $\hat{V}(\hat{p})$

Confidence Intervals

- Estimates are not enough - need indication of their accuracy
- $100(1-\alpha)\%$ Confidence Interval (CI): interval estimate of a population parameter with a certain amount of confidence (NOT prob!)
- Heuristics: If we take repeated samples from the population and construct CIs using this procedure for all the samples, $100(1-\alpha)\%$ of them are expected to capture the true value of the parameter.
- $100\alpha\%$ of all samples generate CIs that do not capture the true value of the parameter
- In practice, we take 1 sample - it's CI either contains or does not contain the true parameter value
- $100(1-\alpha)\%$: confidence level
- In probability sampling with finite populations, can calculate exact confidence level if we are able to generate all possible samples since probability of selection is known theoretically
- We do not always know values of statistics for all possible samples - need other methods practically
- Relied on asymptotics to construct CIs (infinite populations) - use something similar in SRS: finite population is thought of as a subset of a superpopulation which is a subset of another, and so on, ...

Central Limit Theorem of SRS

If n, N, and $N - n$ are sufficiently large* then under certain regularity conditions:

$$\frac{\bar{y} - \bar{y}_U}{\sqrt{(1 - \frac{n}{N})\frac{S}{\sqrt{n}}}} \sim N(0, 1)$$

A large sample $100(1 - \alpha)\%$ CI for the population mean is:

$$\left[\bar{y} - z_{\alpha/2}\sqrt{1 - \frac{n}{N}}\frac{S}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\sqrt{1 - \frac{n}{N}}\frac{S}{\sqrt{n}}\right]$$

If S is unknown, then the CI becomes:

$$\left[\bar{y} - z_{\alpha/2}SE(\bar{y}), \bar{y} + z_{\alpha/2}SE(\bar{y})\right]$$

*Note: "Sufficiently Large" in CLT approximation depends on how close the generating distribution of the y_i is to a Normal distribution

close → small size to be "suff. large"
far → big size to be "suff. large"

Example: $N = 100, n = 10$

Sample 1:
$\bar{y}_U = 9.874$, $\bar{y} = 10.14$
Capital $S = \sqrt{9.296}$

So 95% CI is:
$\bar{y} \pm z_{\frac{\alpha}{2}} \sqrt{1 - \frac{n}{N}} \cdot \frac{S}{\sqrt{n}}$, $\alpha = 0.05$, $z_{0.025} = 1.96$
$10.14 \pm 1.96 \sqrt{1 - \frac{10}{100}} \cdot \frac{\sqrt{9.296}}{\sqrt{10}} = (8.35, 11.93)$ → C.I.

this particular C.I. captures the pop mean $\bar{y}_U$, so this C.I. is GOOD!

```
> mean(mydata)
[1] 9.874
> var(mydata)
[1] 10.03023
```

Find a 95% CI for the population mean for each sample:

(i) Sample 1:

```
> sample1<-sample(mydata,10,replace=F)
> mean(sample1)
[1] 10.14
> var(sample1)
[1] 9.296
```

Example (con'd)...

Check by yourself :

(ii) Sample 2:

```
> sample2<-sample(mydata,10,replace=F)
> mean(sample2)
[1] 7.38
> var(sample2)
[1] 8.237333
```

CI : [5.6924, 9.0676)

BAD!

(iii) Sample 3:

```
> sample3<-sample(mydata,10,replace=F)
> mean(sample3)
[1] 11.08
> var(sample3)
[1] 12.16178
```

CI: (9.0286, 13.1314)

good, but a little bit overestimated!