STA 304 Lecture 1.

- Sampling
  interested in population. take measurements on a sample to make inferences about the population.
- Why?
- Too expensive/time consuming, not obtainable/applicable.
- What makes a good sample?
- a sample that is representative of the population:
  characteristics of interest in the population can be estimated from the sample with a known degree of ~~accuracy~~ accuracy.

Def.

- Observation unit/element: An object on which a measurement is taken
- Variable: The characteristic that is measured on the element.
- Target Population: The complete collection of elements we want to make inference about.
- Sample: A subset of a population
- Sampling unit: A unit that can be selected for a sample.
- Sampling frame: A list or specification of all the sampling units.

Limitations of the Survey.

- possible problems
  ① Undercoverage: not all household are visited on telephone directory (unlisted numbers, those without landlines, etc.)
  ② Non-response: certain people may not answer call or the question.
  ③ Include/adjust for confounding factors such as age, race etc.
  ④ Measurement: people may forget, lie, or misinterpret/count the absences.
  ⑤ Make sure ~~the~~ numbers selected 'randomly'.

- Errors
  - Sampling Errors: statistical errors, due to randomness,
  - non-sampling errors: not due to randomness, but to do with the way in which sample was selected / data was collected.
    - errors of non-observation
      - selection bias: occurs when parts of the target population are not included in the sample population or some units are sampled at a different rate than intended.
      - sample of convenience: units in the sample are selected because they are ~~easier~~ available & easy to access.
    - errors of observation
      - measurement bias: occurs when the measuring instrument tends to differ from the true value in one direction.
        (inaccurate responses from respondents or wrong measurements and/or poor survey design ~~resp~~ from investigators)

Reducing errors
- Reducing Selection Bias:
  Use Probability Sampling Methods - Ch 2.
- Reducing Non-response Bias:
  - Callbacks.
  - rewards & incentives
- Reducing Measurement error:
  - careful questionaire design
  - testing survey equipment
  - training interviewers
  - pretesting survey
  - check for accuracy in respondents data.

Lecture 2

- Advantages of sampling:
  1. Provide reliable information about at less cost:
     - can qualify ~~*~~ sampling error when using probability samples
     - does not destroy population in cases where elements must be destroyed to be measured.
  2. Faster data collection
  3. Estimates often more accurate than those based on census.

Numerical Data Summaries/Statistics

Data: $y_1, \cdots, y_n$

sample mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- measures location
- estimates population mean, μ

sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- measures spread
- estimates population variance σ^2

Parameter:
- usually denote θ
- A characteristic of the population - fixed, unknown

Estimator:
- usually $\hat{\theta}$
- a statistic (function of sample data) used to estimate a parameter.

Estimate:
- Numeric value of an estimator

CI:
- $100(1-\alpha)\%$ of sample generate a CI that covers the
- sample size selected to ensure error of estimation is less than B.

$$P(|\hat{\theta} - \theta| < B) = 1 - \alpha .$$

- Compounds of Events
- Probability. $P(A) = \frac{\text{number of outcomes in } A}{\text{total number of outcomes}}$

Axioms of Prob.

properties of prob. in finite sample spaces:

1. $P(\Omega) = 1$
2. $\forall A,\ 0 \le P(A) \le 1$
3. $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ if A_i are disjoint

& 1. $P(A) + P(A^c) = 1$

2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- Conditional Prob.

$P(A|B) = \frac{P(A \cap B)}{P(B)}$, $P(B) \neq 0$

- Independence

$P(A|B) = P(A)$ & $P(B|A) = P(B)$

$P(A \cap B) = P(A)P(B)$

General Result:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1 | A_2 \cap \cdots \cap A_n) \cdot P(A_2 | A_3 \cap \cdots \cap A_n) \cdots P(A_{n-1} | A_n) P(A_n)$$

Simple Random Sampling

Think N balls in a box labelled $1, 2, \cdots, N-1, N$ and draw n

- with replacement
  - N^n possible samples
  - each sample has $\frac{1}{N^n}$ prob. of being selected.
  - order does not matter
- without.
  - $\binom{N}{n} = \frac{N}{n!(N-n)!}$ possible samples
  - each samples has probability of $\frac{1}{\binom{N}{n}}$ of being selected

- order does not matter
- successive draws are NOT independent

R.V. (discrete/continuous)
Probability Distributions
~~p(x)~~ $p(x)=P(X=x)$
1. $0 \le p(x) \le 1 \quad \forall x$
2. $\sum_x p(x)=1$

Expected Values
- Mean/Expected Value/Expectation : $\mu / \mu_x / E(X)$
  expected: average value of RV over the long run.

$$\mu_X = E(X) = \sum_x x\,p(x)$$

$V(X)=E[(X-\mu_X)^2]=E(X^2)-\mu_X^2=Cov(X,X)$
Variance - spread
$\sigma_X = STD(X)=\sqrt{V(X)}$: standard deviation
$Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]=E(XY)-\mu_{XY}$
Covariance - How much two variances vary together (linear)
$Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$: Correlation - Standardized Covariance

properties:
1. $\forall$ function g, $E[g(X)]=\sum_x g(x)p(x)$
2. $a,b \in \mathbb{R} \quad E(aX+b)=aE(X)+b$
3. X,Y independent, $E(XY)=E(X)E(Y)$
   i.e. $Cov(X,Y)=0$
4. $Cov\left[\sum_{i=1}^{n}(a_iX_i+b_i), \sum_{j=1}^{m}(c_jY_j+d_j)\right]=\sum_{i=1}^{n}\sum_{j=1}^{m} a_i c_j Cov(X_i,Y_j)$
5. $V(X+Y)=V(X)+V(Y)+2Cov(X,Y)$
6. $-1 \le Corr(X,Y) \le 1$

Lecture 3.

Indicator variables

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{o.w.} \end{cases}$$

In sampling, we use this RV:

$$Z_i = I(\text{unit } i \text{ is in sample}) = \begin{cases} 1 & \text{if unit } i \text{ is in the sample} \\ 0 & \text{o.w.} \end{cases}$$

Properties

$$P(Z_i=1) = P(i\text{th unit is in the sample}) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

$$E(Z_i) = \frac{n}{N}$$

For $i \neq j$, $P(Z_i Z_j = 1) = P(i, j \text{ are in the sample}) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$

For $i \neq j$, $E(Z_i Z_j) = \frac{n(n-1)}{N(N-1)}$

$V(Z_i) = \frac{n(N-n)}{N^2}$ $\quad Var(Z_i) = E(Z_i)^2 - (E(Z_i))^2 = E(Z_i) - (E(Z_i))^2$

$$= \frac{n}{N} - \frac{n^2}{N^2} = \frac{n(N-n)}{N^2}$$

For $i \neq j$, $Cov(Z_i, Z_j) = \frac{n(n-N)}{N^2(N-1)}$

Estimation: Aim: Estimate Population Total : $t = \sum_{i=1}^{N} y_i$

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

$\bar{y}_S$ = average value of y's ~~saw~~ in sample S

Sampling distribution : ~~$P(\hat{t}=k) = \sum_S \hat{t}_S P(S) = \sum_k k P(\hat{t}=k)$~~

$$P(\hat{t}=k) = \sum_{S:\hat{t}_S=k} P(S)$$

- expected value : $E(\hat{t}) = \sum_S \hat{t}_S P(S) = \sum_k k P(\hat{t}=k)$
- Bias of estimator : $Bias[\hat{t}] = E(\hat{t}) - t$
  - estimator is called unbiased if $E(\hat{t}) = t$
  - the bias is not the same as selection/measurement bias
- variance : $V(\hat{t}) = \sum_S [\hat{t}_S - E(\hat{t}_S)]^2 P(S)$, called precise if var is small
- Mean Squared Error (MSE) $MSE(\hat{t}) = E[(\hat{t} - t)^2] = V(\hat{t}) + [Bias(\hat{t})]^2$

called accurate if MSE is small

Population Parameters & Estimates

- Population total: $t = \sum_{i=1}^{N} y_i$
  Estimate: $\hat{t} = N\bar{y}_s$

- Population Mean: $\bar{y}_u = \frac{1}{N}\sum_{i=1}^{N} y_i$
  Estimate: $\bar{y}_s = \frac{1}{n}\sum_{i \in s} y_i = \bar{y}$

- Variance of Population values: $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_u)^2$
  estimate with sample variance: $s^2 = \frac{1}{n-1}\sum_{i \in s}^{n}(y_i - \bar{y})^2$

- proportion:

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

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

Properties of $\hat{t}$ & $\bar{y}$ (estimate: population & sample mean)

- $E(\hat{t}) = t$, $E(\bar{y}) = \bar{y}_u$
  Both unbiased estimators.

- $Var(\hat{t}) = N^2(1 - \frac{n}{N})\frac{S^2}{n}$ $\quad 1 - \frac{n}{N}$: finite population correction (fpc)
  $Var(\bar{y}) = \frac{S^2}{n}(1 - \frac{n}{N})$

↓ use unbiased estimator s^2 to estimate S^2

$$\hat{Var}(\hat{t}) = N^2(1 - \frac{n}{N})\frac{s^2}{n}$$

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

- other measures: Standard error (SE): square root of estimated variance of the estimator $SE(\bar{y}) = \sqrt{\frac{s^2}{n}(1 - \frac{n}{N})}$

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

$$CV(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})} \Rightarrow \text{plug in } CV(\bar{y}) = \sqrt{1 - \frac{n}{N}}\frac{S}{\sqrt{n}\bar{y}_u} \text{ for } \bar{y}_u \neq 0$$

$$\hat{CV}(\bar{y}) = \frac{\widehat{SE}(\bar{y})}{\bar{y}} = \sqrt{1-\frac{n}{N}}\,\frac{s}{\sqrt{n}\,\bar{y}} \quad \text{for } \bar{y}_u \neq 0$$

Estimating a proportion

$y_i = I$ (unit i has characteristic)

$$p = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_u$$

$$\hat{p} = \bar{y}$$

$\hat{p}$ is unbiased for p b/c $\bar{y}_u = \bar{y}$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - p)^2 = \frac{N}{N-1}p(1-p)$$

$$V(\hat{p}) = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}$$

$$s^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \hat{p})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

$$\hat{V}(\hat{p}) = \left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$$

CI.

- CI → accuracy, a certain of confidence.
  NOT PROB!

Lecture 4 (part 1 / before midterm)

- Sample size Estimation

e : margin of error.

$$n = \frac{z_{\alpha/2}^2 S^2}{e^2 + \frac{z_{\alpha/2}^2 S^2}{N}} = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{where } n_0 = \left(\frac{z_{\alpha/2} S}{e}\right)^2$$

$Var(y_i) = p(1-p) = p - p^2 \Rightarrow S^2 = \frac{1}{4}$ i.e. $p = \frac{1}{2}$ ~~does~~ (maximum) if nothing is said.

- Sampling Weights

For SRS, weight $w_i = \frac{N}{n}$, each unit represents itself + $\frac{N}{n} - 1$ unsampled units in the population.

~~CI for~~ Large sample CIs for a location – Finite Populations

Use finite population correction (fpc) in variance estimates $(1 - \frac{n}{N})$

For mean:

$$\bar{y} \pm z_{\alpha/2}\sqrt{(1-\frac{n}{N})\frac{S^2}{n}} \quad \text{or} \quad \bar{y} \pm z_{\alpha/2}\sqrt{(1-\frac{n}{N})\frac{s^2}{n}}$$

For total:

$$\hat{t} \pm z_{\alpha/2}\sqrt{(1-\frac{n}{N})\frac{N^2S^2}{n}} \quad \text{or} \quad \hat{t} \pm z_{\alpha/2}\sqrt{(1-\frac{n}{N})\frac{N^2s^2}{n}}$$

For ~~population~~ proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{(1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1}}$$

(Two variables comparison won't be covered in midterm)

When should you use SRS?

- easy to design & analyze, but not for all cases.

Use SRS:

when – Little/no extra info is available about characteristics in the population
– Data users insist on SRS formulas:
averaging sample values
– Main interest is multivariate relationships (regression equation)
for the population: easier to perform and interpret for SRSs.

Do not use SRS:

when – a controlled experiment is appropriate (not a survey sample)
– list of observation units in population is not available or too expensive/time consuming to take SRS
– have additional info about population characteristics that can improve survey design/cost effective design

Find

304 Review (mainly after lecture 4)

For SRS, comparing 2 means: infinite populations

$100(1-\alpha)\%$ Approximate CI for 2 means:

distance of

$$(\hat{\mu}_1 - \hat{\mu}_2) \pm Z_{\alpha/2}\sqrt{V(\hat{\mu}_1) + V(\hat{\mu}_2) - 2Cov(\hat{\mu}_1, \hat{\mu}_2)}$$

- if 0 is ~~the~~ in the interval, then there is no statistically significant difference between μ_1 & μ_2.
-

(finite populations)

2 means:

$$(\bar{y}_1 - \bar{y}_2) \pm Z_{\alpha/2}\sqrt{(1-\frac{n_1}{N_1})\frac{s_1^2}{n_1} + (1-\frac{n_2}{N_2})\frac{s_2^2}{n_2}}$$

2 totals:

$$(\hat{t}_1 - \hat{t}_2) \pm Z_{\alpha/2}\sqrt{(1-\frac{n_1}{N_1})\frac{N_1^2 S_1^2}{n_1} + (1-\frac{n_2}{N_2})\frac{N_2^2 S_2^2}{n_2}}$$

2 proportions:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{(1-\frac{n_1}{N_1})\frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} + (1-\frac{n_2}{N_2})\frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1}}$$

- Stratified Random Sampling (STRS)

Recall ~~RS~~ SRS:

- use SRS when:
  - little/no info. about characteristics in the population.
  - data user insists: averaging sample values.
  - main interest is multivariate relationships (regression equations) ... easier to do in SRS.

- don't use SRS:
  - a controlled experiment is appropriate.
  - no sampling frame
  - additional info ...

- L strata. ~~(AKA~~ mutually exclusive)
- Take SRS indep. in each stratum.
- pool info. to get overall pop. parameters.

- Why STRS?

- Theory & notations for STRS
  - ~~H = 1~~

    $N_1, \cdots, N_L$, $N = \sum_{i=1}^{L} N_i$
  - Take SRS of size n_i from each stratum, denoted S_i.
  - total sample size. $n = \sum_{i=1}^{L} n_i$
  - $i = 1, \cdots, L$: ~~label~~ index for strata.
  - $j = 1, \cdots, N_i$: index for elements within stratum i.

- Population parameters.

  y_{ij} = variable/measurement value of the jth unit in stratum i.

  $\tau_i = \sum_{j=1}^{N_i} y_{ij}$: pop. total in stratum i

  $\tau = \sum_{i=1}^{L} \tau_i$: pop total (overall)

  $\bar{y}_{iU} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$ pop mean in stratum i

  $\bar{y}_U = \frac{\tau}{N} = \frac{\sum_{i=1}^{L} \sum_{j}^{N_i} y_{ij}}{N}$ pop mean

Big S

  $S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{iU})^2$: pop variance within stratum i

  $S^2 = \frac{1}{N-1} \sum_{i=1}^{L} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_U)^2$ pop variance (overall)

- sample quantities/est.

  Use SRS within each stratum:

  $\bar{y}_i = \frac{1}{n_i} \sum_{j \in S_i} y_{ij}$ estimates $\bar{y}_{iU}$

  $\hat{\tau}_i = N_i \bar{y}_i$ estimates τ_i

  $s_i^2 = \frac{1}{n_i - 1} \sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2$ estimates S_i^2

  $\hat{\tau}_{st} = \sum_{i=1}^{L} \hat{\tau}_i = \sum_{i=1}^{L} N_i \bar{y}_i$ estimates τ

  $\bar{y}_{st} = \frac{\hat{\tau}_{st}}{N} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$ estimates $\bar{y}_U$

  weighted average of sample stratum averages.
  weights are prop. of pop. units in each stratum.

Properties of estimators

- unbiased : $\bar{y}_{st}$ is unbiased for $\bar{y}_u$
  $\hat{\tau}_{st}$ is unbiased for τ
- variances: $V(\bar{y}_{st}) = \sum_{i=1}^{L} (1 - \frac{n_i}{N_i})(\frac{N_i}{N})^2 \frac{S_i^2}{n_i}$
  $V(\hat{\tau}_{st}) = \sum_{i=1}^{L} (1 - \frac{n_i}{N_i}) N_i^2 \frac{S_i^2}{n_i}$
- SE : In order to est. var, at least sample 2 units, ow. var = 0.

$$SE(\bar{y}_{st}) = \sqrt{\sum_{i=1}^{L} (1 - \frac{n_i}{N_i})(\frac{N_i}{N})^2 \frac{S_i^2}{n_i}}$$

$$SE(\hat{\tau}_{st}) = \sqrt{\sum_{i=1}^{L} (1 - \frac{n_i}{N_i}) N_i^2 \frac{S_i^2}{n_i}}$$

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

since SRS taken from each stratum i, $E(\bar{y}_i) = \bar{y}_{iu}$

so $E(\bar{y}_{st}) = \sum_{i=1}^{L} \frac{N_i}{N} E(\bar{y}_i) = \frac{1}{N}(\sum_{i=1}^{L} N_i \bar{y}_{iu})$

$$= \frac{\tau}{N}$$

$$= \bar{y}_u$$

unbiased.

Also due to SRS,

$$Var(\bar{y}_i) = (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i}$$

So $Var(\bar{y}_{st}) = Var(\sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i) = \sum_{i=1}^{L} V(\frac{N_i}{N} \bar{y}_i)$ since stratum indep from each other

$$= \sum_{i=1}^{L} \frac{N_i^2}{N^2} V(\bar{y}_i)$$

$$= \sum_{i=1}^{L} (\frac{N_i}{N})^2 (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i}$$

$\hat{\tau}_{st}$ is easy. skipped.

CI ① Sample size with each stratum large OR ② Large number of strata.

$100(1-\alpha)\%$ CI for $\bar{y}_u$ is

$$\bar{y}_{st} \pm z_{\alpha/2} SE(\bar{y}_{st})$$

$100(1-\alpha)\%$ CI for τ is

$$\hat{\tau}_{st} \pm z_{\alpha/2} SE(\hat{\tau}_{st})$$

"just remember p.t ± $z_{\alpha/2}$ SE(p.t)"

STRS for proportions:

$p = \bar{y}_U$

$Y = \begin{cases} 1 & \text{with } p \\ 0 & \text{with } 1-p \end{cases}$

$\hat{p}_i = \bar{y}_i$

$s_i^2 = \frac{n_i}{n_i - 1}\hat{p}_i(1-\hat{p}_i)$

$\hat{p}_{st} = \sum_{i=1}^{L} \frac{N_i}{N}\hat{p}_i$

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

- Sampling Weights:

$\pi_{ij} = \frac{n_i}{N_i}$ => so the sampling weights are $w_{ij} = \frac{1}{\pi_{ij}} = \frac{N_i}{n_i}$

so each sample unit in stratum i represents: itself + $\left(\frac{N_i}{n_i} - 1\right)$ unsampled units in stratum i.

sum of weights is N. i.e. $N = \sum_{i=1}^{L}\sum_{j \in S_i} w_{ij}$

- $\hat{t}_{st} = \sum_{i=1}^{L}\sum_{j \in S_i} w_{ij} y_{ij}$

$\bar{y}_{st} = \frac{\hat{t}_{st}}{N} = \frac{\sum_{i=1}^{L}\sum_{j \in S_i} w_{ij} y_{ij}}{\sum_{i=1}^{L}\sum_{j \in S_i} w_{ij}}$

- STRS is self-weighting if the sampling fraction $\frac{n_i}{N_i}$ is the same for each stratum (i.e. sampling weight is $\frac{N}{n}$ like for SRS.
  But variance depends on stratification - ~~weights~~

Analysis of Variance (ANOVA)

STRS is efficient (low var.) when:

① The observations are homogeneous within strata & heterogenous between strata.

② Strata means differ widely so that the variation among strata is high & the variation with each is small

ANOVA Table for Population

Source	degree of freedom (df)	Sum of Squares
Between Strata	$L-1$	$SSB=\sum_{i=1}^{L}\sum_{j=1}^{N_i}(\bar{y}_{iu}-\bar{y}_u)^2=\sum_{i=1}^{L}N_i(\bar{y}_{iu}-\bar{y}_u)^2$
within Strata	$N-L$	$SSW=\sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\bar{y}_{iu})^2=\sum_{i=1}^{L}(N_i-1)S_i^2$
Total (about $\bar{y}_u$)	$N-1$	$SSTO=\sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\bar{y}_u)^2=(N-1)S^2$

So accordingly, good $\Rightarrow$ SSB ↑, SSW ↓

~~Allocation~~ Ob Allocating Observations to Strata:

- Allocation: How to determine the number of observations to sample in each stratum.

③ factors: ① ~~to~~ total # of elements in each stratum
② variability of observations within each stratum
③ cost of obtaining one observation from each stratum

3 schemes:

① Proportional

- number of sampled units in each stratum is prop. to size of stratum in population.
- prop ensures that sample reflects population wrt stratification variable & sample is a min version of population
- self-weighting sample!
- when strata large enough $V_{prop}(\bar{y}_{st}) \leq V_{SRS}(\bar{y})$ with same sample size. Advantages!

② ~~optim~~ Optimal Allocation

- ... var of estimator is ↓ minimized for a given total cost.
- optimal allocation → smaller cost when S_i^2 differ a lot
  (you want to sample heavily on those with higher variation)

Objective: gain most info. with least cost.:

$$C=C_0+\sum_{i=1}^{L}c_i n_i$$

minimize $V(\bar{y}_{st})$ for a given total C;
minimize C for a fixed $V(\bar{y}_{st})$

$n_i \propto \frac{N_i S_i}{\sqrt{C_i}}$

$\Rightarrow$ ☆ $n_i = \left(\frac{\frac{N_i S_i}{\sqrt{C_i}}}{\sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{C_i}}} \right) n$

- Sample heavily on a stratum if
  - stratum accounts for large part of pop.
  - variance within stratum is large
  - sampling from stratum is expensive

③ Neyman Allocation:
- special case of ② when costs in each strata are (approx) equal.
- $n_i \propto N_i S_i$
  better than prop. allocation!

Sample size:

$n = \frac{z^2_{\alpha/2} V}{e}$ where $V = \sum_{i=1}^{L} \left(\frac{n}{n_i}\right)\left(\frac{N}{N}\right)^2 S_i^{*2}$

where S_i^* is an est. of S_i,
e is margin of error.

Ratio & Regression Estimation

Ratio est. in a ~~RSS~~ SRS

Ratio: $R = \frac{\tau_y}{\tau_x} = \frac{\bar{y}_u}{\bar{x}_u} = \frac{\mu_y}{\mu_x}$

y_i : variable of interest

x_i : auxiliary / subsidiary variable.

population correlation coeff.

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_u)(y_i - \bar{y}_u)}{(N-1) S_x S_y}$$ higher ρ, better est.

Estimators using SRS:

- estimator of pop ~~ratio~~ ratio R:

$$r = \hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{\tau}_y}{\hat{\tau}_x}$$

- estimator of pop total τ_y: $\hat{\tau}_{yr} = r\tau_x$
- estimator of pop mean $\bar{y}_u$: $\hat{\bar{y}}_r = r\bar{x}_u$
- more precise due to high correlation between x & y. (lower var.)
- but var ↓, bias ↑

~~Use MSE to~~ use ratio when we can model x & y with a straight line ~~through origin~~, & var. of y is proportional to x.

One-Stage Cluster Sampling

by location convenient to sample.

psus (primary)
ssus (secondary)

Why? no frame ; occurs in natural clusters ; economical.
When? elements in a cluster heterogeneous, clusters are homogeneous. } opposite of STRS

Population Quantities at psu level.

N = number of clusters/psus in the population

M_i = number of ssus in psu i, $i = 1, \ldots, N$

$M = \sum_{i=1}^{N} M_i$ = total # of ssus in the pop.

$\bar{M} = \frac{M}{N}$ = average cluster size for the pop.

y_{ij} = measurement for jth element in psu i.

$T_i = y_i = \sum_{j=1}^{M_i} y_{ij}$ = total in psu i

$T = \sum_{i=1}^{N} T_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$ = population total

Population Quantities at ssu level

$\bar{y}_u = \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$ = population mean

$\bar{y}_{iu} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \frac{T_i}{M_i}$ = pop mean in psu i

$S^2 = \frac{1}{M-1} \sum_{i=1}^{N} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_u)^2$ = pop var. (per ssu)

$S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{iu})^2$ = pop var within psu i.

Sample Quantities

n = number of psus in the sample

m_i = number of ssus in the sample from psu i, $i = 1, 2, \cdots, N$

$m_i = M_i$ for one-stage cluster sampling

S = sample of psus

S_i = sample of ssus from ith psu

$\bar{y}_i = \frac{1}{m_i} \sum_{j \in S_i} y_{ij}$ = sample mean for psu i

$\hat{T}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} = y_i$ estimated total for psu i

$\bar{y}_t = \frac{1}{n} \sum_{i \in S} y_i$ = average of the sampled cluster totals

$s_t^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y}_t)^2$ = sample variance of psu totals

$s_i^2 = \frac{1}{m_i - 1} \sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2$ = sample var within psu i.

Ratio Estimation

Ratio Estimator of Population Mean, $\bar{y}_u$: $\bar{y} = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} m_i}$

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

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

if M is unknown, use $\bar{m}$ to estimate $\bar{M}$

if M is known, use Ratio Est. for population Total, τ:

$\hat{\tau}_r = M\bar{y}$ with $SE(\hat{\tau}_r) = MSE(\bar{y})$

Unbiased Estimation

$$\hat{\tau}_{unb} = N\bar{y}_t = \sum_{i \in s} \frac{N}{n} y_i$$

$$\hat{V}(\hat{\tau}_{unb}) = N^2\left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}, \quad s_t^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_t)^2$$

cluster of equal size

$M_i = m_i = m$

$M = Nm$, total sample size mn

overall average of all nm sample measurements

$$\bar{\bar{y}}_c = \frac{1}{mn} \sum_{i \in s} \sum_{j=1}^{m} y_{ij}$$

$$\bar{y}_t = m\bar{\bar{y}}_c$$

$$\hat{V}(\bar{\bar{y}}_c) = \left(1 - \frac{n}{N}\right) \frac{1}{nm^2} s_t^2 = \left(1 - \frac{n}{N}\right) \sum_{i \in s} (\bar{y}_i - \bar{\bar{y}}_c)^2$$

& now the ratio est. & unb. est. are the same.

Sampling weights. (one-stage)

$$w_{ij} = \frac{1}{P(\text{ssu } j \text{ of psu } i \text{ is in sample})} = \frac{N}{n}$$ self-weighting

Theory for clusters of Equal Size.

Intraclass Correlation Coefficient (ICC)
measures homogeneity with clusters.

$$ICC = 1 - \frac{m}{m-1}\frac{SSW}{SSTO}$$

- ICC = 1 when clusters are perfectly homogeneous i.e. SSW = 0

$$MSB = \frac{SSB}{N-1} = \frac{Nm-1}{m(N-1)} S^2[1 + (m-1)ICC]$$

- ICC > 0 if MSB > MSW:
  when elements within a psu are similar & so cluster sampling is less efficient than SRS
- ICC ≤ 0 if MSB ≤ MSW. when elements with a psu are dispersed, cluster means will be similar & so cluster sampling more efficient. than SRS

ICC is for equal cluster sizes

another measure is adjusted R^2, $R_a^2 = 1 - \frac{MSW}{S^2}$

- clusters homogeneous, SSB ↑, SSW ↓ → R_a^2 ↑

Unequal sizes clusters

Ratio est. more efficient.

ANOVA Source	df	SS	MS	
Between psus	$N-1$	$SSB = \sum_{i=1}^{N}\sum_{j=1}^{m}(\bar{y}_{iu} - \bar{y}_u)^2$	$MSB = \frac{SSB}{N-1}$	unbiased
Within psus	$N(m-1)$	$SSW = \sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{iu})^2$	$MSW = \frac{SSW}{N(m-1)}$	unbiased
Total (about $\bar{y}_u$)	$Nm-1$	$SSTO = \sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \bar{y}_u)^2$	~~MSTO~~ $= S^2$	biased

(Two-stage cluster sampling)

when the cluster inside is homogeneous, we might not select all ssus in it. so do another SRS.

Sampling weight. $w_{ij} = \frac{1}{P(\text{ssu } j \text{ of psu } i \text{ is in sample})} = \frac{NM_i}{n m_i}$

Est. pop mean:

~~① M is known~~ ~~$\frac{1}{M}$~~ ~~$\frac{N}{M}$~~

① M is known, unb

② M is unknown, ratio

Est. pop total.

Unb.

Variance from one-stage cluster + additional var due to selection of ssus within psus.

Advantages, Disadvantages of cluster sampling.

- large.
- easy
- opposite to STRS
- one-stage is special (take it all!) $M_i = m_i$
- larger variance than SRS for the same sample size
- more precision per dollar
- unbiased est. v.s. ratio est.
  - if cluster sizes vary greatly, ratio is better (small var)
  - for equal, both are equivalent.

- 2-stage when elements in clusters are homogeneous.

Systematic Sampling

Sampling interval $k = \frac{N}{n}$

Special case of ~~one~~ two? -stage cluster.

k psus of size n

Take SRS of 1 psu.

Est Population mean:

$$\hat{\bar{y}}_{sys} = \bar{y}_i = \bar{y}_{iu}$$

$$E(\hat{\bar{y}}_{sys}) = \bar{y}_u$$

$$V(\hat{\bar{y}}_{sys}) = \frac{S^2}{n}[1+(n-1)ICC]$$

Est Pop total:

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

$$E(\hat{t}_{sys}) = t$$

$$V(\hat{t}_{sys}) = N^2\frac{S^2}{n}[1+(n-1)ICC]$$

$ICC < 0 \Rightarrow$ sys more precise.

ICC large $\Rightarrow$ SRS more precise.

Types of sampling frames

① sampling frame is random.

$ICC \approx 0$.

Behaves like SRS: use SRS results & formula to.

~~Est~~ $\hat{V}(\hat{\bar{y}}_{sys}) = (1-\frac{n}{N})\frac{s^2}{n}$

$$\hat{V}(\hat{t}_{sys}) = N^2(1-\frac{n}{N})\frac{s^2}{n}$$

② increasing/decreasing:

positive autocorrelation:

· $ICC < 0$: $variance_{sys} <$ Var_{SRS} b/c sys forces sample to spread out.

· use SRS but it will overestimate Var.

③ periodic.

underestimate var if use SRS.

systematic sampling less precise than SRS.

Pro/cons.

- get a representative sample but frame not constructed.
- treated as SRS mostly
- each sys. sample is a /psu : use cluster formula to est. var.

pros: . cheaper/easier than SRS/STRS
- provide more info. per unit cost than SRS. when ~~at~~ there is a pattern

Cons: • biased/non-representative if periodic pop.
- Variance under/over estimated if periodic/ordered.

UTM 2012 Test 2 V1

1.
(a). F
(b). T
(c). T
(d). F
(e). T

2. (a). STRS VS Cluster.
① STRS: sample from each strata so groups will be chosen & so elements within are homogeneous & between strata are heterogeneous.

② Cluster: opposite.

(b).

$$\hat{B} = \frac{\bar{y}}{\bar{x}} \qquad \tilde{B} = \frac{\bar{Y}}{\bar{x}_U}$$

we use $\hat{B}$ b/c it increases precision: since x & y are correlated, so $\hat{B}$ has smaller var!

3. (a) 1-stage cluster
(b) 100
(c) 20 23 34 39 69
(d) 67.31

se = ?

(e). Sampling weight for jth student in ith class is $w_{ij} = \frac{N}{n} = 2$

(f). $68.68 = \frac{6731}{5} \times 10 \times \frac{1}{196}$

$$se = \sqrt{\frac{1}{5 \times 19.6^2}\left(1 - \frac{5}{10}\right)\frac{1}{5-1}\sum_{ies}(M_i\bar{y}_i - M\bar{y}_{umb})^2} = 1.105 \text{ ?}$$

$32.39998 + -429.60002 + 373.4000$ $\quad -.106$ $\quad .0998$

B. (a). CO

(b). $\frac{25}{133} = 0.18$ 80

(i).

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

$$= \sum \left(\frac{133}{196} \times 61.88722 + \frac{63}{196} \times 77.15873 \right)$$

$$= 66.79592$$

$$66.79592 \pm 1.96 \sqrt{\left(1 - \frac{25}{133}\right)\left(\frac{133}{196}\right)^2 \frac{268.4767}{25}}$$

$$= (62.3684, 71.2235)$$

V2.

(a). F X T — 1-stage cluster self-weighting

(b). T X F

(c). T

(d). T

(e). F

Overview

STRS

N: pop size

L: strata #

N_i sampling units in stratum i

$N = \sum_{i=1}^{L} N_i$

$n = \sum_{i=1}^{L} n_i$

~~y_{ij}~~

~~$\tau_i = \sum_{j=1}^{N_i} y_{ij}$ population total in stratum i~~

Cluster

Pratice Exam.

1. (a) F (e) T (i) F
   (b) F (f) T (j) T
   (c) T (g) F
   (d) F (h) F

2.

5.

(a) $k = \frac{N}{n} = \frac{100}{20} = 5$

(b) $\hat{\bar{y}}_{sys} = \bar{y} = 9.535$

(c) $ICC = 1 - \frac{n}{n-1}\frac{SSW}{SSTO}$

$= 1 - \frac{20}{19}\frac{3356.3}{3356.3+7.6} = -0.0503$

$V(\hat{\bar{y}}_{sys}) = \frac{S^2}{n}[1 + [illegible](n-1)ICC]$

$= 0.07812$

(d) using SRS,

$(1 - \frac{80}{100})\frac{S^2}{20} = 1.4107$

(e)

6. (a) area.

(b) ratio, highly correlated.

(c) ~~F=2.11E~~ 7850467.29.

Practice Exam

1. (a). F (b). F (c). T (d). F

(e). T (f). T (g). F (h). F

(i). F (j). T

2. (a). V. psus = church
ssus = member
$N = 400$
$n = 50$
M unknown. M_i unknown
$m_i = 20$ $\forall$ all i
Cluster because homogeneous between churches
easy to access.

(b). ~~III~~, I, ~~II~~
have sampling frame.
$N = 10000$
$n = 200$

(c). III,
Better to stratify since the results varies by gender / health condition, ...

(d). X, Reason: Don't know what is related so better to experiment.

(e) VI. Highly correlated. So can do ratio

~~$N = 252$~~
~~$N = 40$~~

y_i = sales of book
x_i = size of bookstore
Appropriate to fit through origin
x & y are pos. correlated
$N = 252$, $n = 40$.

3. (a). STRS, easy to stratify, and the result varies among different previous year's inventories.

(b). $n_1 = \frac{400}{2000} \times 100 = 20$

$n_2 = \frac{1000}{2000} \times 100 = 50$

$n_3 = \frac{600}{2000} \times 100 = 30$ proportional allocation

(c). $\bar{y}_{str} = \sum_{i=1}^{L} N_i \bar{y}_i = 400 \times 105 + 1000 \times 180 + 600 \times 282 = 391200$

$$se = \sqrt{\sum_{i=1}^{L} (1 - \frac{n_i}{N_i}) N_i^2 \frac{S_i^2}{n_i}}$$

$$= \sqrt{(1 - \frac{20}{400}) 400^2 \frac{2400}{20} + (1 - \frac{50}{1000}) 1000^2 \frac{9800}{50} + (1 - \frac{30}{600}) 600^2 \frac{2500}{30}}$$

$= 6974.2383$

$\bar{y}_{str} \pm 1.96 \times 5374.0115 = 391200 \pm 6974.2383$

$(377530.49, 404869.51)$

(d). divide CI above by 2000.

(e). DO SRS. to III.

$600, 282, 2500$

$$\frac{282}{1} \pm 1.96 \sqrt{(1 - \frac{30}{600}) \frac{2500}{30}}$$

$$282 \pm \frac{49\sqrt{14}}{30}$$

$(264.56, 299.44)$

(f) $C_3 = 4C_1$

$C_2 = 2C_1$

$n_1 = \frac{400 \times 20}{\sqrt{C_1}}$

$n_2 = \frac{1000 \times 30}{\sqrt{2C_1}}$

$n_3 = \frac{600 \times 50}{\sqrt{4C_1}}$

$n_1 : n_2 : n_3 = \frac{8000}{\sqrt{C_1}} : \frac{30000}{\sqrt{C_1} \cdot \sqrt{2}} : \frac{30000}{2\sqrt{C_1}}$

$= 8000 : 21213 : 15000$

$n_1 \approx 18$

$n_2 \approx 48$

$n_3 \approx 34$

$n_1 + n_2 + n_3 = 100$ ✓

(g) Yes. B/c strata is large.
Yes. B/c variance varies.

4.
(a) 2-stage cluster
(b) algebra classes.
students
(c) 13
(d) 5 12 24 28 19 17 21 22 3 3 3
(e) 38, 44, 46, 58, 62
(f) 60.49235

$se(\bar{y}) = \sqrt{\frac{1}{nM^2}\left(1-\frac{n}{N}\right)s_1^2 + \frac{1}{nNM^2}\sum_{i \in S} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}}$

$= \sqrt{\frac{1}{5}21.8^2\left(1-\frac{5}{12}\right) \times \frac{1}{4} + \frac{1}{5 \times 12 \times 21.8^2} \times 2281630.7}$

$= \sqrt{225297.1}$

(g). 62.5686

(h). [redacted]

$\binom{19}{3} = 969$

$\frac{\binom{19}{2}}{\binom{19}{3}} = 0.1765 \times \frac{5}{12} = 0.0735$

(i). $\frac{5}{12} = 0.4167$

(j). $\hat{\bar{y}}_{unb} = \frac{N}{n}\sum_{i \in S} M_i \bar{y}_i \Big/ M = \frac{1}{299} \times \frac{12}{5}(21.8 \times 5 \times 60.4924)$

$= 52.9258$

$se(\hat{\bar{y}}_{unb})$

$= \sqrt{\frac{1}{5 \times 21.8^2}\left(1 - \frac{5}{12}\right)\frac{1}{4}\sum_{i \in S}(M_i \bar{y}_i - M \cdot 52.9258)^2 + \frac{1}{5 \times 12 \times 21.8^2} 225297.1}$

$= \sqrt{45.8839}$

$\pm$ [redacted]

[redacted]

(k). 66. $\left(1 - \frac{3}{17}\right) \times \frac{867}{3} = 238$

Supplementary Exercise #3

1. STRS

(a) ~~Show $\hat{\tau}_{str}$ is unbiased.~~

~~$\bar{y}_U = \frac{\tau}{N} = \frac{\sum_{i=1}^{L}\sum_{j=1}^{N_i} y_{ij}}{N}$~~

~~$\hat{\bar{y}}_{str} = \frac{\hat{\tau}_{str}}{N} = \frac{\sum_{i=1}^{L} N_i \bar{y}_i}{N} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$~~

~~$E(\hat{\bar{y}}_{str}) = \bar{y}_U$~~

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

since SRS from each stratum i, $E(\bar{y}_i) = \bar{y}_{iU}$

so $E(\bar{y}_{str}) = \sum_{i=1}^{L} \frac{N_i}{N} E(\bar{y}_i) = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_{iU} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_{iU} = \frac{\tau}{N} = \bar{y}_U$

& $Var(\bar{y}_i) = (1 - \frac{n_i}{N_i})\frac{S_i^2}{n_i}$

$Var(\bar{y}_{str}) = Var\left(\sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i\right) = \sum_{i=1}^{L} Var\left(\frac{N_i}{N} \bar{y}_i\right)$ since each stratum indep from each other

~~$Var(\hat{\bar{y}}_{str}) = Var($~~

$= \sum_{i=1}^{L} \frac{N_i^2}{N^2} Var(\bar{y}_i)$

$= \sum_{i=1}^{L} \frac{N_i^2}{N^2} (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i}$

Since $\hat{\tau}_{str} = N\bar{y}_{str} \Rightarrow E(\hat{\tau}_{str}) = NE(\bar{y}_{str}) = N\bar{y}_U = \tau$

$Var(\hat{\tau}_{str}) = Var(N\bar{y}_{str}) = N^2 Var(\bar{y}_{str})$

$= \sum_{i=1}^{L} N_i^2 (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i}$

2. b/c $\hat{p}_{str} = \sum_{i=1}^{L} \frac{N_i}{N} \hat{p}_i$

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

$= \bar{y}_{str}$ same

3. $w_{ij} = \frac{1}{P(j\text{th sample in } i\text{th stratum is in sample})} = \frac{1}{\frac{n_i}{N_i}} = \frac{N_i}{n_i}$

self-weighting $\Rightarrow \frac{n_i}{N_i}$ same for all i

$\Rightarrow \bar{y}_{str} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i = \sum_{i=1}^{L} \frac{N_i}{N} \cdot \frac{1}{n_i} \sum_{j \in S_i} y_{ij}$

$= \frac{1}{N} \sum_{i=1}^{L} \sum_{j \in S_i} \frac{N_i}{n_i} y_{ij}$

4. 2 strata.

$S_2^2 = k S_1^2$ $\quad \frac{S_1}{S_2} = \sqrt{\frac{1}{k}}$

$n_i \propto N_i S_i$

$\frac{n_1}{N_1 S_1} = \frac{n_2}{N_2 S_2}$

(1) $\frac{n_1}{n_2} = \frac{N_1 S_1}{N_2 S_2} \Rightarrow \frac{N_1}{N_2} \sqrt{\frac{1}{k}}$

(2) proportional?

$k = 1 \Rightarrow \frac{n_1}{N_1} = \frac{n_2}{N_2}$

Supplementary 4

1. 4 reasons why ratio est. ~~give ex~~ & give ex.

① To estimate a ratio
② To estimate a population total when N is unknown
③ To increase precision of estimates (when the two RVs are highly correlated)
④ To adjust estimates to reflect demographic totals.
⑤ To adjust for Non-response.

2. Why we estimate $\bar{x}_U$ with $\bar{x}$ in ratio estimator even we assume we know the population quantities for x.

③ when Ratio instead of SRS/regression?
when x & y are perfectly correlated. then they are the same.

But most appropriate when we can model the relationship between x & y with a straight line through origin & var of y is proportional to x.

4. ① $r = \frac{\tau_a}{\tau_1} = \frac{30}{21} = \frac{10}{7}$

$\hat{\tau}_{1r} = \frac{\tau_{2sr}}{r} = \frac{7526}{\frac{10}{7}} = 5268.2 \doteq 5268$ households

② $r = \frac{80}{21}$

$\frac{80}{21} \times 5268 = 20068.57 = 20069$ ~~people~~ people

③ $\frac{32}{21} = 1.524$ TV/household

④ $\frac{31}{21} \times 5268 = 7776.5714$ cars

⑤ $\frac{16}{21} \times 5268 = 4013.7143$ households

⑥ MSE of total cars?

⑦ 6342 households.

(i). SRS $\frac{31}{21} \times 6342 = 9362$ cars

(ii). $(6342)^2 (1 - \frac{21}{6342}) \frac{S^2}{21}$ = ...

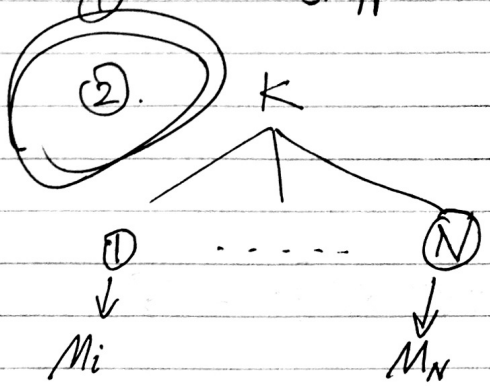
when ~~$S^2 = \frac{1}{6342}$~~ $S^2 = \frac{1}{20}(Y_j - \frac{37}{21})^2$

(iii). MSE?

(iv). $MSE(\hat{\bar{Y}}_r) \leq MSE(\bar{y})$ iff $\rho > \frac{rS_x}{2S_y} = \frac{CV(x)}{2CV(y)}$.

Supplementary 5

① --- skipped.



$SSTO = (K-1)S^2$

$SSTR = \sum_{i=1}^{N} M_i^2 (\bar{y}_{iu} - \bar{y}_u)^2$

$SSE = \sum_{i=1}^{N} (M_i - 1) S_i^2 =$ ~~$\sum_{i=1}^{N} (M_i - 1) \frac{1}{N-1}$~~

~~$SSTR + SSE = \sum_{i=1}^{N} \left(M_i^2 (\bar{y}_{iu} - \bar{y}_u)^2 + M_i S_i^2 - S_i^2 \right)$~~

~~$= \sum_{i=1}^{N}$~~

~~$SSTR = K \sum_{j=1}^{K} (\bar{y}_i - \bar{y}_u)^2$~~

~~$= K(K-1)S^2$~~

~~$SSE = \sum_{i=1}^{N} (M_i - 1) \frac{1}{M_i - 1} SSW = NSSW$~~

~~$SSE = \sum_{i=1}^{N} (M_i - 1) \cdot \frac{1}{M_i - 1} SSW = \sum_{i=1}^{N} SSW$~~

~~$SSTR = \sum_{i=1}^{N} M_i^2$~~

3. $G_1 = \{4, 5, 8\}$

$G_2 = \{2, 6, 8\}$

(a). $\bar{y}_1 = \frac{4+5+8}{3} = \frac{17}{3}$

$\bar{y}_2 = \frac{2+6+8}{3} = \frac{16}{3}$

For variance 要不要减一

$Var_1 = \left[\left(\frac{17}{3} - 4\right)^2 + \left(\frac{17}{3} - 5\right)^2 + \left(\frac{17}{3} - 8\right)^2\right] \frac{1}{3} = \frac{26}{9}$

$Var_2 = \frac{56}{9}$

(b). $\bar{y}_U = \left(\frac{17}{3} + \frac{16}{3}\right) \times \frac{1}{2} = \frac{11}{2}$

$\left[\left(\frac{11}{2} - \frac{17}{3}\right)^2 + \left(\frac{11}{2} - \frac{16}{3}\right)^2\right] \frac{1}{2} = \frac{1}{36}$

(c). $ICC = 1 - \frac{m}{m-1} \frac{SSW}{SSTO}$

$=$

$m = 3$

$SSW = \sum_{i=1}^{2} \sum_{j=1}^{3} (y_{ij} - \bar{y}_{iU})^2$

$= \frac{26}{3} + \frac{56}{3} = \frac{82}{3}$

$SSTO = (Nm-1)S^2 = (6-1) \cdot (5.5 - y_{ij})^2$

$= 5 \times \frac{55}{2}$

$= 137.5$

Furthermore

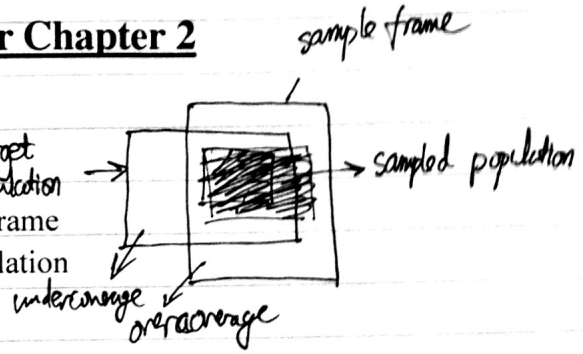
So SRS

S. $ICC = 1 - \frac{3}{2} \cdot \frac{82/3}{157.5} = 0.7018$

(d). $ICC > 0$ elements in psu similar, SRS better than Cluster

STA304H1F - Summer 2014:
Surveys, Sampling, and Observational Data

SUPPLEMENTARY EXERCISES for Chapter 2



1. a) Give a specific example in which:

both can be represented by "telephone survey"

(i) An element is not in the target population but in the sampling frame

(ii) An element is not in the sampling frame but in the target population

b) What is the statistical term for the bias in (i) ? overcoverage

c) What is the statistical term for the bias in (ii)? undercoverage

2. A survey is conducted to determine how much beer is consumed by U of T students. A simple random sample of registered students is asked "*How much beer do you drink* ?"

Briefly explain what is wrong with this question and reword it to improve the question.

The ~~used~~ question is vague, the ~~quit~~ quantity of beer is hard to calculate. And this question is without a time scale...

3. True/False?

a) The sampled population is always a subset of the target population. ~~False~~ False

b) Selection bias is a type of non-sampling error. True

c) A census will obtain more accurate results than a sample. F

d) Giving rewards and incentives to participants of surveys will increase response accuracy. rate? True False

4. Refer to the "Hite Report" from Lohr's textbook-(excerpt is given on the next page) to answer the following questions:

a) How were the respondents 'self-selected' ? Why should this be avoided in surveys?

b) Why might one expect these respondents to be more educated than the general population?

c) Give an example of a leading question in this particular survey.

d) Identify the following:

(i) Target population (ii) Sampling frame (iii) Observational unit (iv) Sampling unit

Excerpt from Lohr , 2 ed (Pg#1):

Shere Hite's book *Women and Love: A Cultural Revolution in Progress* (1987) had a number of widely quoted results:

- 84% of women are "not satisfied emotionally with their relationships" (p. 804).
- 70% of all women "married five or more years are having sex outside of their marriages" (p. 856).
- 95% of women "report forms of emotional and psychological harassment from men with whom they are in love relationships" (p. 810).
- 84% of women report forms of condescension from the men in their love relationships (p. 809).

The book was widely criticized in newspaper and magazine articles throughout the United States. The *Time* magazine cover story "Back Off, Buddy" (October 12, 1987), for example, called the conclusions of Hite's study "dubious" and "of limited value."

Why was Hite's study so roundly criticized? Was it wrong for Hite to report the quotes from women who feel that the men in their lives refuse to treat them as equals, who perhaps have never been given the chance to speak out before? Was it wrong to report the percentages of these women who are unhappy in their relationships with men?
Of course not. Hite's research allowed women to discuss how they viewed their experiences, and reflected the richness of these women's experience in a way that a multiple choice questionnaire could not. Hite erred in generalizing these results to all women, whether they participated in the survey or not, and in claiming that the percentages above applied to all women. The following characteristics of the survey make it unsuitable for generalizing the results to all women.

- The sample was self-selected—that is, recipients of questionnaires decided whether they would be in the sample or not. Hite mailed 100,000 questionnaires; of these, 4.5% were returned.
- The questionnaires were mailed to such organizations as professional women's groups, counseling centers, church societies, and senior citizens' centers. The members may differ in political views, but many have joined an "all-women" group, and their viewpoints may differ from other women in the United States.
- The survey has 127 essay questions, and most of the questions have several parts. Who will tend to return such a survey?
- Many of the questions are vague, using words such as "love." The concept of love probably has as many interpretations as there are people, making it impossible to attach a single interpretation to any statistic purporting to state how many women are "in love." Such question wording works well for eliciting the rich individual vignettes that comprise most of the book, but makes interpreting percentages difficult.
- Many of the questions are leading—they suggest to the respondent which response she should make. For instance: "Does your husband/lover see you as an equal? Or are there times when he seems to treat you as an inferior? Leave you out of the decisions? Act superior?" (p. 795)

Hite writes "Does research that is not based on a probability or random sample give one the right to generalize from the results of the study to the population at large? If a study is large enough and the sample broad enough, and if one generalizes carefully, yes" (p. 778). Most survey statisticians would answer Hite's question with a resounding "no." In Hite's survey, because the women sent questionnaires were purposefully chosen and an extremely small percentage of those women returned the questionnaires, statistics calculated from these data cannot be used to indicate attitudes of all women in the United States. The final sample is not *representative* of women in the United States, and the statistics can only be used to describe women who would have responded to the survey.

Hite claims that results from the sample could be generalized because characteristics such as the age, educational, and occupational profiles of women in the sample matched those for the population of women in the United States. But the women inthe sample differed on one important aspect—they were willing to take the time to fill out a long questionnaire dealing with harassment by men, and to provide intensely personal information to a researcher. We would expect that in every age group and socioeconomic class, women who choose to report such information would in general have had different experiences than women who choose not to participate in the survey.

1. (b)

Known $E(\bar{y}) = \bar{y}_u$, $E(\hat{t}) = t$

Show $V(\bar{y})$ & $V(\hat{t})$

poly. expansion

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

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

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

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

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

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

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

$$= \frac{1}{n}(1-\frac{n}{N})S^2$$

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

Similarly, for proportion:

- $p = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}_u$, $y_i = I$ (z_i)
- $\hat{p} = \bar{y}$ is unbiased for $\bar{y}_u = p$, done.

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_u)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - p)^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - 2p\sum_{i=1}^{N} y_i + Np^2\right)$$

$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i - 2p\sum_{i=1}^{N} y_i + Np^2\right)$$

$$= \frac{1}{N-1}\left((1-2p)\sum_{i=1}^{N} y_i + Np^2\right)$$

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

$$= \frac{1}{N-1}\left[pN - 2p^2N + p^2N\right)$$

$$= \frac{1}{N-1}\left[pN - p^2N\right]$$

$$= \frac{1}{N-1} pN(1-p)$$

$$= \frac{N}{N-1} p(1-p)$$

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

$$= \frac{N}{N-1} p(1-p)\frac{1}{n}(1-\frac{n}{N})$$

$$= \frac{N-n}{N-1}\frac{p(1-p)}{n}$$

STA304/1003 H1F - Summer 2014:
Surveys, Sampling, and Observational Data

Supplementary Exercises # 2

1. Let $\mathcal{S}$ be a SRS of size n from a population of size N and let

$$Z_i = \begin{cases} 1, & i \in \mathcal{S} \\ 0, & i \notin \mathcal{S} \end{cases}$$

(a) How many samples contain i ? $\binom{N-1}{n-1} = \frac{(N-1)!}{(n-1)!(N-1-n+1)!}$

(b) Find $P(Z_i = 1)$. $\frac{n}{N}$

(c) How many samples contain both i and j (where $i \neq j$)? $\binom{N-2}{n-2}$

(d) Find the probability that both i and j are in $\mathcal{S}$. $\frac{n(n-1)}{N(N-1)}$

(e) Derive $E(Z_i)$, $V(Z_i)$, and $Cov(Z_i, Z_j)$ for $i \neq j$. $E(Z_i) = p(Z_i=1)\cdot 1 + p(Z_i=0)\cdot 0 = P(Z_i=1) = \frac{n}{N}$

$V(Z_i) = E(Z_i^2) - E(Z_i)^2 = E(Z_i) - E(Z_i)^2 = \frac{n}{N} - \frac{n^2}{N^2} = \frac{n(N-n)}{N^2}$

$Cov(Z_i, Z_j) = E(Z_i Z_j) - E(Z_i)E(Z_j)$

$= \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N}$

$= \frac{n(n-1)N - n^2(N-1)}{N^2(N-1)}$

$= \frac{n^2N - nN - n^2N + n^2}{N^2(N-1)}$

$= \frac{n(n-N)}{N^2(N-1)}$

(f) Are Z_i and Z_j independent? Explain. Find the distribution of Z_i. NO! Bernoulli. $(\frac{n}{N})$

(g) Show that $\bar{y}$ is unbiased for $\bar{y}_U$ and $\hat{t}$ is unbiased for t .

$E(\bar{y}) = E[\frac{1}{n}\sum_{i\in\mathcal{S}} y_i] = \frac{1}{n}E(\sum_{i=1}^{N} Z_i y_i) = \frac{1}{n}\sum_{n=1}^{N} y_i E(Z_i) = \frac{1}{n}\sum_{i=1}^{N}\frac{n}{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U$; $E(\hat{t}) = NE(\bar{y}) = N\bar{y}_U = t$

(h) Derive $V(\bar{y})$ and $V(\hat{t})$.

$V(\bar{y}) = V(\frac{1}{n}\sum_{i\in\mathcal{S}} y_i) = \frac{1}{n^2}V(\sum_{i=1}^{N} Z_i y_i) = \frac{1}{n}(1-\frac{n}{N})S^2$

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

(i) Show that s^2 is unbiased for S^2.

$s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{N} Z_i(y_i - \bar{y})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2 = S^2$

2. Consider the following example in which one wishes to take a sample of size $n = 2$ from population with $N = 8$:

$\binom{8}{2} = \frac{8!}{2!6!} = \frac{7\times 8}{1\times 2} = 28$ possible samples

i	1	2	3	4	5	6	7	8
y_i	1	2	8	2	8	1	5	1

$\sum_{i\in\mathcal{S}} y_i$	2	3	4	6	7	9	10	13	16
# of samples	3	6	1	3	2	6	4	2	1

(a) Find the sampling distribution of $\hat{t}$. Sampling distribution of $\hat{t} = N\bar{y}_S$

(b) Find the sampling distribution of the $\bar{y}$. $\frac{\sum_{i\in\mathcal{S}} y_i}{2}$

$E(\hat{t}) = \sum_k k P(\hat{t}=k) = \dots$ equivalent

$t = \sum y_i =$

(c) Use the sampling distributions to verify:

~~(i)~~ the unbiasedness of $\hat{t}$ and $\bar{y}$

(ii) formulae for the variances of $\hat{t}$ and $\bar{y}$ as derived in 1.(h) above

$V(\hat{t}) = E(\hat{t}^2) - [E(\hat{t})]^2 = \dots$

3. Consider the population of undergraduate students who are currently registered in the Statistics program at University of Toronto. Let

$$y_i = \begin{cases} 1, & \text{if student i wishes to pursue graduate studies in Statistics} \\ 0, & \text{otherwise} \end{cases}$$

(a) Give an expression for the true proportion of students in this population who wish to pursue graduate studies in Statistics. $p = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U$

(b) ~~Prove that for 0-1 data such as this, the following is true:~~

$$S^2 = \frac{N}{N-1}\bar{y}_U(1-\bar{y}_U)$$

$\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2$

$= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2$

$= \frac{1}{N-1}\sum_{i=1}^{N}(y_i^2 - 2y_i\bar{y}_U + \bar{y}_U^2)$

(c) Does a similar formula hold for s^2?

Yes $s^2 = \frac{n}{n-1}\sum_{i\in s}(y_i - \hat{p})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$ where $\hat{p} = p = \bar{y}_U$

(d) Suppose the population has 4000 students and we take a SRS of 200. 125 students say they wish to pursue graduate studies in Statistics.

(i) Find a 95% CI for the true proportion of students in this population who wish to pursue graduate studies in Statistics.

$\frac{125}{200} \pm 1.96\sqrt{(1-\frac{200}{4000})\frac{\frac{125}{200}(1-\frac{125}{200})}{199}} = (0.5594, 0.6906)$

(ii) Find a 95% CI for the population proportion of students who do not wish to pursue graduate studies in Statistics. (0.3094, 0.4406)

(iii) Find a 95% CI for the population percentage of students who wish to pursue graduate studies in Statistics. (55.94, 69.06)

(iv) Find a 95% CI for the population total number of students who wish to pursue graduate studies in Statistics. take $S = \frac{1}{2}$!

$\hat{t} = 4000 \times \frac{125}{200} = 2500$

$2500 \pm 1.96\sqrt{(1-\frac{200}{4000})\frac{4000^2 \cdot \frac{1}{2}^2}{200}} = (2365, 2635)$

4. Give an example of a scenario for sampling in which each unit has equal probability of selection but it is not a SRS.

When we need to study different groups of the target population...

Supplementary Exercise #1 Recap.

1.

(a). $\binom{N-1}{n-1}$

(b). $P(Z_i=1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{(N-1)!}{(n-1)!(N-n)!} \cdot \frac{n!(N-n)!}{N!} = \frac{(N-1)!\,n!}{(n-1)!\,N!}$

$= \frac{n}{N}$

(c). $\binom{N-2}{n-2}$

(d). $P(Z_i=1 \,\&\, Z_j=1) = \frac{n}{N} \cdot \frac{(n-1)}{N-1} = \frac{n(n-1)}{N(N-1)}$

(e) $E(Z_i) = P(Z_i=1)\cdot 1 + P(Z_i=0)\cdot 0 = \frac{n}{N}$

$V(Z_i) = E(Z_i^2) - (E(Z_i))^2$

$= E(Z_i) - (E(Z_i))^2$

$= \frac{n}{N} - \frac{n^2}{N^2}$

$= \frac{n(N-n)}{N^2}$

$Cov(Z_i, Z_j) = E(Z_i Z_j) - E(Z_i)(Z_j) = \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N} = \frac{Nn(n-1) - n^2(N-1)}{N^2(N-1)}$

$= \frac{Nn^2 - Nn - n^2N + n^2}{N^2(N-1)}$

$= \frac{n(n-N)}{N^2(N-1)}$

(f). Z_i & Z_j not independent, $Z_i \sim$ Bernoulli (p) where $p = \frac{n}{N}$

(g). Show $\bar{y}$ unbiased for $\bar{y}_u$ & $\hat{t}$ unbiased for t.

$E(\bar{y}) = E(\frac{1}{n}\sum_{i\in s} y_i) = E(\frac{1}{n}\sum_{i=1}^{N} Z_i y_i) = \frac{1}{n}\sum_{i=1}^{N} y_i E(Z_i) = \frac{1}{n}\cdot\frac{n}{N}\sum_{i=1}^{N} y_i = \bar{y}_u$

$E(\hat{t}) = E(N\cdot\bar{y}) = NE(\bar{y}) = N\cdot\bar{y}_u = t$

(b) $V(\bar{y}) = V(\frac{1}{n}\sum_{i\in S} y_i) = \frac{1}{n^2} V(\sum_{i=1}^{N} z_i y_i)$

$= \frac{1}{n^2}[\sum_{i=1}^{N} y_i^2 V(z_i) - \sum_{i\neq j} y_i y_j \, Cov(z_i z_j)]$

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

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

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

$= \frac{N-n}{nN^2}\frac{1}{N-1}[N\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} y_i^2 - ((\sum y_i)^2 - \sum y_i^2)]$

$= \frac{N-n}{nN^2}\frac{N}{N-1}[\sum_{i=1}^{N} y_i^2 - \frac{(\sum y_i)^2}{N}]$

$= \frac{N-n}{nN}\cdot\frac{1}{N-1}[\sum_{i=1}^{N} y_i^2 - \frac{(\sum y_i)^2}{N}]$

$= \frac{N-n}{nN}\cdot S^2$

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

(i) Show $E(s^2) = S^2$? See next page

$= \frac{1}{n-1}E(\sum_{i\in S}(y_i^2 - 2y_i\bar{y} + \bar{y}^2))$

~~Show s^2 is unbiased for S^2~~

~~$E(s^2) = E\left[\frac{1}{n-1}\sum_{i\in S}(y_i - \bar{y})^2\right]$~~

~~$= \frac{1}{n-1} E\left[\sum_{i\in S}[(y_i - \bar{y}_u) - (\bar{y} - \bar{y}_u)]^2\right]$~~

~~$= \frac{1}{n-1} E\left[\sum (y_i - \bar{y}_u)^2 - n(\bar{y} - \bar{y}_u)^2\right]$~~

~~$= \frac{1}{n-1}\left[\sum E(y_i - \bar{y}_u)^2 - n E(\bar{y} - \bar{y}_u)^2\right]$~~

~~$= \frac{1}{n-1}\left[\quad\right] S^2$~~

Show s^2 is unbiased for S^2
(sample variance) (pop variance)

$(n-1)E(s^2) = E\left[\sum_{i=1}^{n} y_i^2 - 2\bar{y}\sum y_i + \bar{y}^2\right]$

$(n-1)E(s^2) = E\left[\sum_{i=1}^{n} y_i^2\right] - E\left[2\bar{y}\sum y_i\right] + E\left[\bar{y}^2 \sum 1\right]$

$= E[\sum y_i^2] - E[2\bar{y}(n\bar{y})] + E[\bar{y}^2 \Sigma 1]$

$= E[\sum y_i^2] - 2nE[\bar{y}^2] + nE[\bar{y}^2]$

$= nE[y_i^2] - nE[\bar{y}^2]$

$\frac{n-1}{n} E(s^2) = E[y_i^2] - E[\bar{y}^2]$

Let $Y = \bar{y}$

$E[Y^2] = E[\bar{y}^2] = Var[Y] + (E[Y])^2$

$= var[\frac{1}{n}\Sigma y_i] + \bar{y}_u^2$

$= \frac{1}{n^2} Var(\Sigma y_i) + \bar{y}_u^2$

$= \frac{1}{n^2}\Sigma Var(y_i) + \bar{y}_u^2$

$= \frac{1}{n^2}\Sigma S^2 + \bar{y}_u^2$

$= \frac{1}{n} S^2 + \bar{y}_u^2$

$\frac{n-1}{n} E(s^2) = E[y_i^2] - E[\bar{y}^2]$

$= S^2 + \bar{y}_u^2 - \frac{1}{n}S^2 - \bar{y}_u^2 = (1 - \frac{1}{n}) S^2$

so $E(s^2) = S^2$ done

(proportion)

For ~~binary~~. show $E(s^2) = S^2$.

$$E(s^2) = E\left(\frac{n}{n-1}\hat{p}(1-\hat{p})\right)$$

$$= \frac{n}{n-1} E(\hat{p}(1-\hat{p}))$$

$$= \frac{n}{n-1}\left[E(\hat{p}) - E(\hat{p}^2)\right]$$

$$= \frac{n}{n-1}\left[p - (V(\hat{p}) + E(\hat{p})^2)\right]$$

$$= \frac{n}{n-1}\left[p - \frac{N-n}{N-1}\frac{p(1-p)}{n} - p^2\right]$$

$$= \frac{n}{n-1}\left[p(1-p) - \frac{N-n}{N-1}\frac{p(1-p)}{n}\right]$$

$$= \frac{n}{n-1}p(1-p)\left[1 - \frac{N-n}{N-1}\frac{1}{n}\right]$$

$$= p(1-p)\frac{n}{n-1}\left[\frac{(N-1)n - N + n}{(N-1)n}\right]$$

$$= p(1-p)\frac{n}{n-1}\frac{N(n-1)}{(N-1)n}$$

$$= \frac{N}{N-1}p(1-p) \quad \checkmark$$

$$= S^2$$

UTM 2012 Fall Test 1 V1.

1. (a). T (b). F (c). T (d). F (e). ~~T~~ F

2. (a). True. ~~$P(S)=\frac{1}{5}$, $\pi_i=5$~~

False. A SRS in this case would mean each ~~student~~ sample of size 8 would have the same prob of selection. But $P(4,4)=0$, $P(5,3)\neq 0$. So NO

(b). ~~The test marks in~~ STA304 students

list ~~marks~~ of 25 Sta & 15 non-sta attending this week's group study.

a ~~each~~ student's ~~mark~~

the selected 8 students' marks

(c). undercoverage. — some people never attends the group study

overcoverage — students of the selected 8 might miss the test.

(d). ~~due to random selection~~. it occurs because it's a sample, not a census, different samples yield diff. results & estimates. (inevitable)

(e). take a SRS from entire class or Stratified sample.

3. (a). 120 25

(b). 71.2 , 66

(c). 70.27187

(d). ~~69.908~~ ~~$\sqrt{\ }$~~

70.27187

$E(\bar{y})=\bar{y}_U$

~~$s=\frac{1}{25-1}\sum_{i\in S}(y_i-69.908)^2$~~

~~$\sqrt{4.8916}=2.2117$~~

sample estimated variance $(1-\frac{25}{120})\cdot\frac{s^2}{25}$, $s^2=5.8511$

$SE=\sqrt{\text{plug in}}$

?

0.3936

(e). expected value of sample variance is 5.8511 (b/c it's unbiased)

(f). $\bar{y}\pm 1.96\sqrt{(1-\frac{25}{120})\frac{5.8511}{25}} = [69.4282, 71.1155]$

70.27187

~~$\bar{y}\pm Z_{\alpha/2}\sqrt{(1-\frac{25}{120})\frac{4.8916}{25}}$~~

Why sample mean, & sample variance.

4. (a). ~

(b). $P(i,j \text{ both selected}) = \frac{1}{\binom{10}{2}} = \frac{1}{\frac{10!}{2!8!}} = \frac{1}{45}$

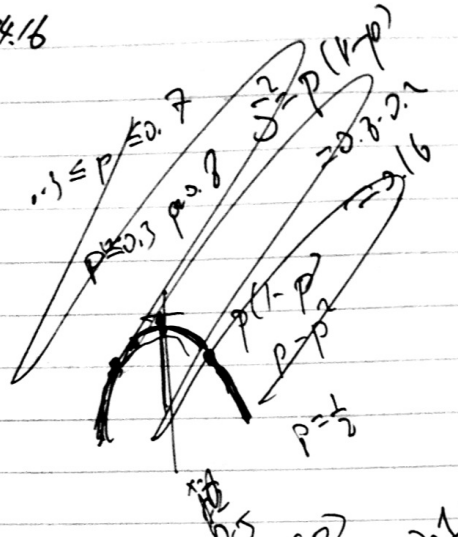
(c). ~~48~~ 9

(d) 0

5.

(a). $n = \frac{n_0}{1+\frac{n_0}{N}}$ where $n_0 = \left(\frac{Z_{\alpha/2} S^*}{e}\right)^2 = \left(\frac{1.96 \times \frac{1}{2}}{0.05}\right)^2 = 384.16$

$$n = \frac{384.16}{1+\frac{384.16}{580}} = 232$$



(b) $\hat{p} = \frac{120-93}{120} = 0.225$

$$\hat{p} \pm 1.96\sqrt{(1-\tfrac{120}{580})\frac{\hat{p}(1-\hat{p})}{119}}$$

$$= 0.225 \pm 1.96\sqrt{\tfrac{7}{9} \times \frac{0.225 \times 0.775}{119}}$$

$= (0.1588, 0.2912)$ not that high (4%)

Ber. (p)

$E(y) = P(y=1)\cdot 1 + P(y=0)\cdot 0$

$= p \cdot 1 + (1-p)\cdot 0$

$= p$

$S^2 = Var(y) = E(y^2) - E(y)^2$

$= p - p^2$

$= p(1-p)$

$E(\hat{p}) = \bar{y}$

$V(\hat{p}) = E(\hat{p}^2) - E(\hat{p})^2 = \bar{y} - \bar{y}^2 =$

$S^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$

bc $V(\hat{p}) = E(\hat{p}^2) - E(\hat{p})^2$

want to show $E(S^2) = S^2$

$$E(S^2) = E\left(\frac{n}{n-1}\hat{p}(1-\hat{p})\right) = \frac{n}{n-1}E(\hat{p} - \hat{p}^2) = \frac{n}{n-1}\left[E(\hat{p}) - V(\hat{p}) - E(\hat{p})^2\right]$$

$$= \frac{n}{n-1}\left[p - \frac{N-n}{N-1}\frac{p(1-p)}{n} - p^2\right]$$

$$= \frac{n}{n-1}\, p(1-p)\left(1 - \frac{N-n}{(N-1)n}\right)$$

$$= \left(\frac{n}{n-1}\right)\left(\frac{(N-1)n - N + n}{(N-1)n}\right) p(1-p)$$

$$= \frac{N}{N-1}\, p(1-p)$$

$$= S^2$$

✓

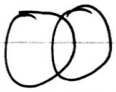
UTM 2012 Fall Test 1 V2

1. (a). F (b). T (c). F (d). F (e). T

UTSG 2008/2009 Test 1.

1. (a). All shoppers at a national chain store.

(b). "cluster sampling"

(c). No. 

(d). money spent /hr /person /per

• satisfaction

(e). Short questionaire.

2. skipped.

3. 2, 4, 6, 10.

$\bar{y} = \frac{5+4+3+3}{4} = \frac{15}{4} = 3.75$

$S^2 = \cdots \quad \sigma^2 = \frac{N-1}{N} S^2 = \frac{11}{12} 0.917 = 0.840$ (unbiased)