W2 Mere Randomization

2.1 Simple Random Sampling

Frame & RN's
• Sample size n from population size N
• Using random numbers from a table, match numbers to unique numbers assigned to each frame element
  - for each selection after the 1st, check to see if the frame element has already been selected
  - if selected already, reject the selection and use another random number
  - continue process until n distinct frame element selected
• Without replacement selection: unique sample elements
• Every element of the population has the same probability of selection (epsem = equal probability selection method) and every combination of size n has the same probability of selection

Selecting a Sample
• Sample size n from population size N
• Using random numbers from a table, or a statistical software system, assign a random number to each of the N frame elements
  - sort the frame elements by random number, from smallest to largest
  - select the first n frame elements
• This is without replacement & epsem
• Random number generation for n = 500:
    RN = TRUNC(URAN(0718)*500)+1
    - URAN: starting point e.g. 0718 ➔ start from page 7 column 18
    - TRUN: truncate off the decimal point

Selecting a sample – Another method
• sample size n from population size N
• using random numbers from a table, match numbers to unique numbers assigned to each frame element
  - continue process until n frame element selected
  - check for duplicates in sample
  - If any duplicates occur, reject sample, and draw another sample of size n
  - without replacement & epsem
  - restricted (simple) RS v.s. unrestricted RS

Definitions of simple random sampling
• Any procedure with fixed sample size n and for which every element of the population has the same probability of selection (epsem) and every combination of size n has the same probability of selection
• All sets of size n distinct elements from N – pick one (N choose n)

Practical Use
• Widely used for simple problems
• But rarely used by practitioners in 'isolation'
  - complicated for 'lay' administration
  - more efficient methods available
  - relies only on randomization
• For practitioner a tool to be used in conjunction with other methods
  - random sample of elements within a group
  - random sample of groups

2.2. A short history

• Sampling practice
  - result of attempts to solve practical problems
• Function of theory
  - formalize implicit assumptions, and confirm, correct or extend practice
• Origins
  - data gathering
    - health and social problems
    - social physics
  - census

- monography

Representative Method
• Kaier: Representative method
  - miniature of country
  - large number of units
  - use prior information in selection
• Von Mayr and others: Census
  - no calculation where observation is possible
• Cheysson and others: Monography
  - detailed examination of typical cases

Randomization
• Representative
  - purposive sampling
  - expert choice
  - balanced sampling
• Objective
  - randomized selection
  - Bowley, 1906 (colleague of R.A. Fisher)
• Neyman 1934
  - The sampling distribution
    - properties of sample under repeated sampling: All possible samples and their associated probabilities of occurrence
    - the sampling distribution of an estimator

Comparison
• Conditions under which different procedures will produce valid estimates
  - probability sampling
    - "unbiased" irrespective of population structure
  - purposive/balanced/quota sampling
    - tough assumptions about population structure, unlikely to be achieved in practice

Principles
• Probability sampling for objectivity
• Stratification for precision (representativeness)
• Variance estimation from the sample
• Complete and comprehensible description of the sampling procedure

2.3 SRS sampling distributions

Basic Framework
• A sample design for which the unit of selection is population element
• Basic framework: Neyman 19334
  - must be application to all populations
  - must not depend on assumptions about the population structure
  - appropriate for large populations of elements
• Repeated sampling
  - objective (mechanical) selection of elements
  - consider possible outcomes of the sampling process
  - evaluation of the whole set of possible outcomes
• The set of all possible values of the estimator that can be obtained with a given sample design
  - for a given sample we obtain a particular value, the estimate (such as y_bar)
• We want to know …
  - … how likely is the estimate to be close to the population value?
• In fact, we select just one sample
• The estimate may be correct, or incorrect
• Want to maximize the probability of a satisfactory estimate

Properties of the sampling distribution
• Unbiasedness
- expected value (average value): E(y_bar)
- meaning of expected value:

$$E(\bar{y}) = \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \bar{y}_s$$

• **Meaning of unbiasedness:**

$$E(\bar{y}) = \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \bar{y}_s = \bar{Y}$$

• standard error
For our SRS of n = 20,

$$var(\bar{y}) = \frac{(1-f)}{n} s^2$$
$$= \frac{\left(1 - \frac{20}{370}\right)}{20} 766.62$$
$$= 36.26$$
$$se(\bar{y}) = \sqrt{\text{var}(\bar{y})} = 6.02$$

• Variability from one sample to another
- variance of the estimator: Var(y_bar)
- meaning of the variance:

$$Var(\bar{y}) = \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \left(\bar{y}_s - E(\bar{y})\right)^2$$

- Algebraically equivalent formula:

$$Var(\bar{y}) = \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \left(\bar{y}_s - E(\bar{y})\right)^2 = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

• Variability from one sample to another
- components

S^2

$\left(1 - \frac{n}{N}\right) = (1-f)$ fpc

$\frac{1}{n}$

- scale conversion

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

• Estimating variability from one sample to another
- element variance: S^2
- estimated element variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \left(= \frac{n}{n-1} p(1-p)\right)$$

- estimated variance & standard error

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

PRECISION

Confidence Intervals
• For large samples, the sampling distribution of y_bar is normal
 - law of large numbers or central limit theorem
• Form an interval around y_bar:

$$\bar{y} \pm 1.96 \times se(\bar{y})$$

• (1-alpha)% or 95% confidence interval
• A statement of uncertainty about our estimated mean

2.4 Sample Size

What we need to know
• What sample size do we need to obtain a give standard error of the estimator?
• S^2 population variance known (or guessed)
 - census
 - other surveys
 - administrative records
• Desired standard error
 - policy requirements in terms of root(Var(y_bar))
 - decision making requirements

Sample Size Formula

• From previous lecture, $Var(\bar{y}) = \left(1-\frac{n}{N}\right)\frac{S^2}{n}$

• For an infinitely large population (or for sampling with replacement), this is

$$Var(\bar{y}) = \frac{S^2}{n}$$

• We can calculate the necessary sample size to achieve desired variance $V_d = Var(\bar{y})$ as

$$n = S^2/V_d$$

• Let's call n' the necessary sample size –

$$n' = \frac{S^2}{V_d}$$

• To calculate the actual n needed for a population of a particular size, we adjust --

$$n = \frac{n'}{1+\frac{n'}{N}}$$

• In general (that is, not assuming N is large), the variance may be expressed as

$$Var(\bar{y}) = \left(1-\frac{n}{N}\right)\frac{S^2}{n} = \frac{S^2}{n'}$$

where

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

Example
• Interested in U.S. population attitudes about how well its current president is doing his or her job
 - "Do you approve or disapprove of the job President Obama is doing as President?" (if approve/disapprove, ask:) "Do you approve/ disapprove strongly or somewhat?"
• Estimate the proportion P approving strongly or somewhat in a new survey
• Suppose p = 0.6 in the last survey
 - then s^2 = p(1-p) = 0.6(1-0.6) = 0.24
• for our new survey about to be conducted, "project" that S^2 = 0.24
• Also need to specify precision of the new survey estimate … in advance … the V_d = Var(y_bar)
• Suppose we would like to end up with an uncertainty statement that says that between 58% and 62% of the U.S. population think President Obama is doing a good job … at a 95% level of confidence
• Recall that the upper confidence limit, the 62% value, is the proportion, 60% in this case, plus a multiplier times the standard error

• That is, 62% = 60% + z x se(60%)
• For a 95% confidence interval, z = 1.96, say z = 2
• Then, if 62% = 60% + 2 x se(60%), se(60%) = 1%
• If that's the kind of confidence interval we want, then we want a standard error of 1%
• of course, se(p) = 0.01 is another way to say this, in terms of what we want to have happened
• proportions are better to work with than percentages
• We need the square of the standard error, or the variance
V_d = Var(p) = (SE(p))^2
That is V_d = (0.01)^2 = 0.0001
• Hence, we have S^2 = 0.24 and V_d = 0.0001
• This yields a necessary sample size of:

$$n' = \frac{S^2}{V_d} = \frac{0.24}{0.0001} = 2{,}400$$

• **Adjustment for the finite population:**

$$n = \frac{n'}{1+\frac{n'}{N}} = \frac{2{,}400}{1+\frac{2{,}400}{250{,}000{,}000}} = 2{,}399.97 = 2{,}400$$

2.5 Margin on Error

Two questions to consider following up from the previous section…
• Is there a more direct way to figure this out from a projected confidence interval?
• Why doesn't the population size have a big effect on the sample size?

Using Desired standard errors
• Recall that we got the necessary sample size n' from:
n'=S^2 / V_d
• And then we could calculate the actual n needed for a population of a particular size by:

$$n = \frac{n'}{1+\frac{n'}{N}}$$

• The example developed a desired level of precision from the width of a confidence interval:
(Lower limit, Upper Limit) = (p – z x se(p), p + z x se(p))
• We set upper and lower limits for a 95% confidence interval, where z = 2 (approximately – 1.96 exactly for large samples):
(Lower 95% limit, Upper 95% limit) = (p – 2 x se(p), p + 2 x se(p))
• Suppose we want
(Lower 95% limit, Upper 95% limit) = (0.58, 0.62)
• Then some will refer to the **margin of error e** as the distance from the upper limit to the middle, or the lower limit to the middle
- in most practice margin of error is about proportions or percentages, as here
• In some areas of application of probability sampling, this distance is referred to as the "**precision**"
• Calculate then
E = 2 x se(p) = (U-L) / 2 = (0.62-0.58) / 2 = 0.02
• In a newspaper report, you might see then the "margin on error" reported, but never the standard error…
- President Obama's approval rating now stands at 60% (plus or minus 2%)
• The public has gotten used to forming the 95% confidence interval from this statement
• It's only one step to get the desired standard error and sampling variance:
Root(V_d) = e/2 = 0.02/2 = 0.01
V_d = 0.0001
• But some trained are to use e directly in calculating sample size
- you may see sample size formulas that are based on e
- these alternative formulas yield the same result as what we do here
- but it can be confusing, especially if one has learned one way rather than the other
• The necessary sample size formula using e is:

$$n' = \frac{S^2}{\left(\frac{e}{2}\right)^2}$$

or **4S^2/e^2** or **(z^2 * S^2) / e^2** or **z^2 * p * (1-p) / e^2** or $\frac{z^2_{1-\alpha/2}\,p(1-p)}{e^2}$

• And this can be then 'adjusted' to obtain the final sample size as

$$n = \frac{n'}{1+\frac{n'}{N}}$$

• And finally, the calculation can also be done in one step, rather than two:

$$n = \frac{S^2}{\left(\frac{e}{2}\right)^2 + \frac{S^2}{N}}$$

• For our example, then where e=0.02,

$$n = \frac{0.24}{\left(\frac{0.02}{2}\right)^2 + \frac{0.24}{250{,}000{,}000}} = 2{,}399.97 = 2{,}400$$

2.6 Sample and population size

• Suppose we are evaluating presidential approval or leadership approval across a number of countries
• We are not sure what the approval rating will be in each
 - use p = 0.50 or the largest value of S^2 = p(1-p) possible
 - this may specify a sample size larger than needed in communities where p is not 0.50
 - in the absence of more precise information about the approval in a community, use the 'conservative' value of p=0.50 and S^2 = 0.25
• What sample size is needed in china with N = 800,000,000 if for a 95% confidence interval e=0.02

$$V_d = \left(\frac{e}{2}\right)^2 = \left(\frac{0.02}{2}\right)^2 = 0.01^2 = 0.0001?$$

Calculate

$$n = \frac{S^2}{\left(\frac{e}{2}\right)^2 + \frac{S^2}{N}} = \frac{0.25}{\left(\frac{0.02}{2}\right)^2 + \frac{0.25}{800{,}000{,}000}} = 2{,}499.99 = 2{,}500$$

• what about approval in the U.S. with N = 250,000,000 and 95% confidence interval with e = 0.02?

$$n = \frac{S^2}{\left(\frac{e}{2}\right)^2 + \frac{S^2}{N}} = \frac{0.25}{\left(\frac{0.02}{2}\right)^2 + \frac{0.25}{250{,}000{,}000}} = 2{,}499.97 = 2{,}500$$

$$n = \frac{S^2}{\left(\frac{e}{2}\right)^2 + \frac{S^2}{N}} = \frac{0.25}{\left(\frac{0.02}{2}\right)^2 + \frac{0.25}{80{,}000}} = 2{,}424.24 = 2{,}425$$

• What about approval in the Seychelles with N = 80,000?
• What about approval in Tuvalu with N = 8,000?

$$n = \frac{S^2}{\left(\frac{e}{2}\right)^2 + \frac{S^2}{N}} = \frac{0.25}{\left(\frac{0.02}{2}\right)^2 + \frac{0.25}{8{,}000}} = 1{,}904.76 = 1{,}905$$

Sample size depends on population size, but not in an expected way

It is clearly not proportional:

China	N = 800,000,000	n = 2,500
USA	N = 250,000,000	n = 2,500
Ireland	N = 4,000,000	n = 2,500
Seychelles	N = 80,000	n = 2,424
Tuvalu	N = 8,000	n = 1,904

• Why do we bring this up? Cuz there are textbooks out there that claim sample size should a fraction of the population, say 10%
 - thus the larger the population, the larger the sample size
 - directly proportional
• This is a common sense misperception
 - How can a sample of only 800 represent the voting public of 250,000,000 in the U.S.?
• The constant fraction sample size (f = n / N) clearly misleads:
 - studies can't have a relatively small sample size to get any useful results for a large country …