W1 Sampling as a research tool

1.1 Why Sample at all?

Research Design and Sampling
• Experiments
  - dependent variable
  - factors
  - control or randomization of disturbing variables

• 1954 Field of the Salk Polio Vaccine
  - two study designs: Observed control & double blind randomized control experiment
    - 220,000 vaccinated 2nd graders & 725,000 unvaccinated 1st & 3rd graders
    - 200,000 vaccinated 2nd graders & 20,000 controls
    - conclusion from randomized study: vaccine effective, safe

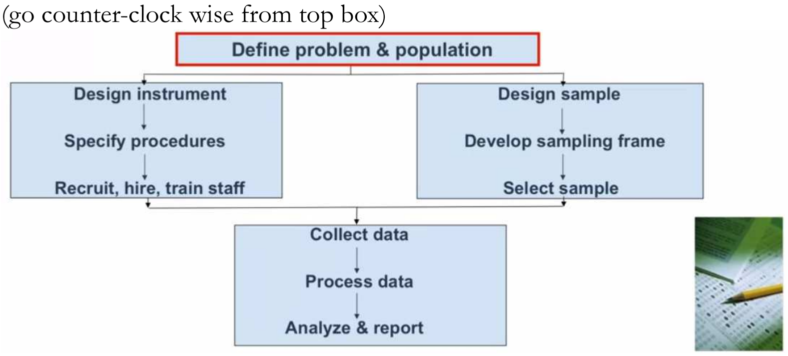
• Doll Hill 1951-1954 British Doctors Study (more of a pseudo-experiment)
  - Survey of all registered physicians in UK
  - 40,000 (2/3) responded & followed – no randomization
  - link between smoking & lung cancer, heart disease

• Quasi Experimental: observational
• Survey Samples: Observational

3Rs
• Realism
• Randomization
• Representation: Randomization doesn't always guarantee representation

1.2 Is a sample a just a sample, or are there types of sampling available?

Problem ➔ Measurement ➔ results ➔ Sampling
(go counter-clock wise from top box)

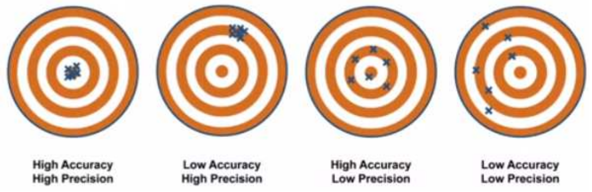


1.3 Why Sample?

Census or Sample
• During conceptualization, a researcher considers the <u>relevant population</u> for evaluating the theory/hypothesis
• In designing the data collection, the researcher has two concerns in mind:
  - external validity
  - cost/benefit calculations for the overall cost of the study
• census involves an enumeration of a population. When the population is large:
  - it is costly
  - it is time consuming
  - it may not be feasible with complete precision (US census as an example)
• A **sample** involves a selection of a <u>representative</u> subset of a **population** in order to draw inferences to the population
• Collecting data from a sample of a large population is **FAR LESS costly and FAR LESS time consuming**

• How do samples get collected?
- recruitment directly – **volunteer** samples
- lists, selection & then recruitment
- lists, selection, recruitment & **non-response**

Accuracy
• Because of the cost savings, sampling allows a researcher to devote
- more resources to the collection of more data (variables)
- the reduction of error in measurement (reliability and validity)
- better coverage of the units of analysis
• This fits in with what is called a <u>Total Survey Error</u> Perspective



Probabilities
• Non-Probability sampling
- haphazard, convenience, or accidental sampling
- purposive sampling or expert choice
- quota sampling: e.g. interview 10 in this town, 4 being from this race and 4 being from this gender etc.
- substitution (for non response)
- online panels
- river sampling
• Probability sampling
- simple random selection
- stratified selection
- cluster samples
- systematic samples
- more complex samples: probabilities proportionate to size

Frames
• List frame
• Area Frame
• Problems
- missing elements
- duplicate listings
- clusters
- blanks or ineligibles

Techniques
• simple random sampling
• systematic sampling
• stratified sampling
- proportionate allocation
- disproportionate allocation
• Cluster sampling
• Two-stage sampling
• Probability proportionate to size sampling
• Stratified probability proportionate to size sampling
• Multistage sampling
• Multiple phase sampling

Deficiencies
• Nonresponse
- total/unit
- item
• Noncoverage

• Compensation: weighting
  - unequal probabilities
  - nonresponse
  - non-coverage (post-stratification)
    - make the sample distribution conform to known population distribution

Complex Design
• Complex designs typically involve one or more of …
  - stratification
  - clusters
  - weights
• Estimation becomes complex
  - even a simple mean or proportion requires non-standard techniques
• standard software cannot handle complex sample designs correctly
• estimating precision becomes more complex as well
• methods of variance estimation must be considered
  - taylor series approximation
  - balanced or jackknife repeated replication
• computer software available for these methods
  - requires stratum, cluster and weight on each sample record

1.4 Why Randomize

Why might we randomize and how might we do it?

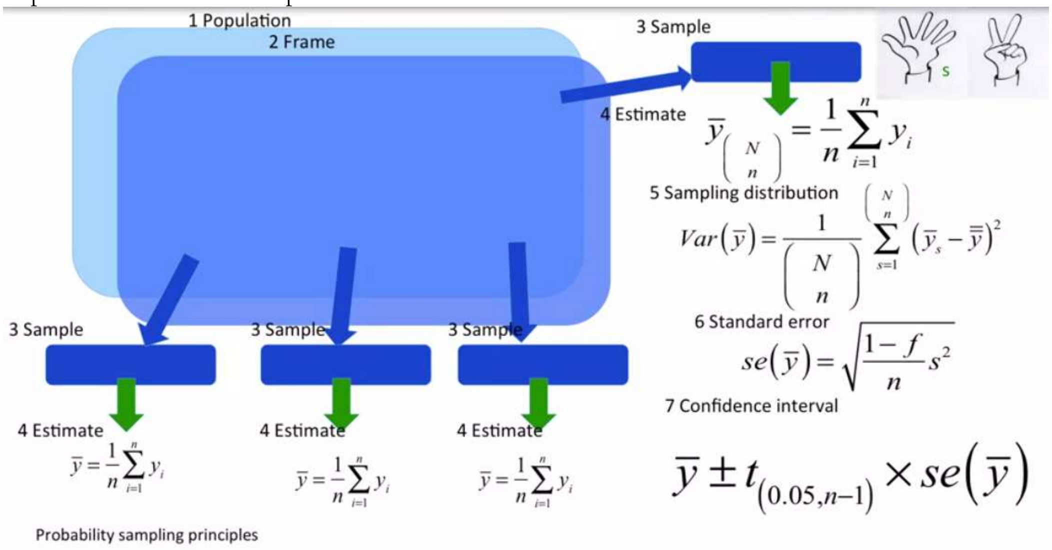
Random Numbers
• 10 random numbers
  - From the Uniform Distribution
• A string of 50 random numbers: 34042253511835630477……
  - also from the uniform distribution
• The string of random 50 numbers in 10 blocks of five each
• 10 random numbers from a normal distribution – more numbers concentrated in the middle

Use in sample selection
• sample selection: frame – n = 370
• Numbering:
  - 8 digit ID
  - sequence no.
• Match random numbers (they come first) to list

Should we put it back
• What if we get the same random number more than once in a sample?
  - keep it: with replacement selection (WR)
  - drop it: without replacement selection (WOR)
• Preference: drop it – better results

1.5 What happens when we randomize?

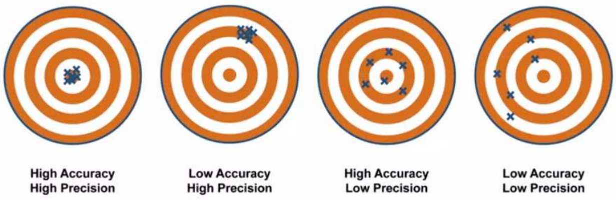
Population to frame to sample



• Thus, two measures…
  - bias
  - variance
• And a random process
  - using random digits applied to a frame to generate, in theory, a large number of possible samples
  - And we can measure the variance across all possible samples from a single randomly drawn sample
  - But only random samples allow us to do this without making any assumptions about either …
    - the sampling mechanism
    - the population distribution

1.6 How do we evaluate how good the sample is?

• Standard error of P decreases and precision increases as sample size increases
• Two measures of data quality
  - bias: we can determine theoretically if a sampling technique is unbiased
  - variance (standard error) – we can determine from sample data alone the size of the variance … to compare numerically



1.7 What kinds of things can we sample?

People / Records / Networks