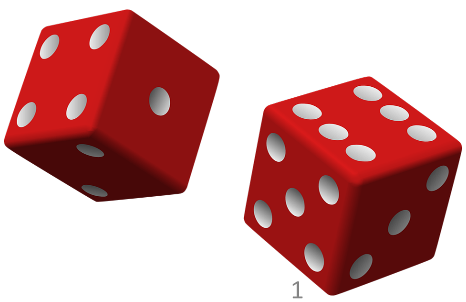
Sampling People, Records, & Networks

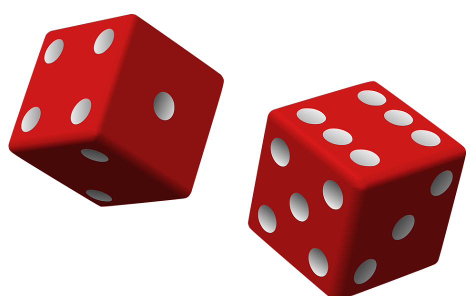
Jim Lepkowski, PhD
Professor & Research Professor *Emeritus*
Institute for Social Research, University of Michigan
Research Professor,
Joint Program in Survey Methodology, University of Maryland



Unit 3

- 1 Simple complex
- 2 deff & roh
- 3 2-stage sampling
- 4 Designing 2-stage samples
- 5 Unequal sized clusters
- 6 Subsampling

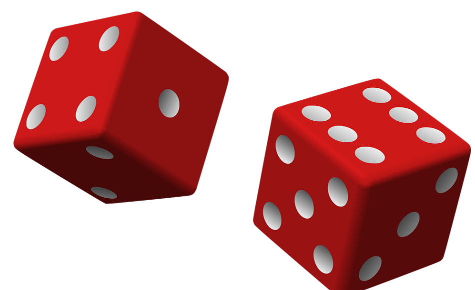
- Unit 1: Sampling as a research tool
- Unit 2: Mere randomization
- Unit 3: Saving money
  - Lecture 1: Simple complex sampling – choosing entire clusters
  - Lecture 2: Design effects & intraclass correlation
  - Lecture 3: Two-stage sampling
  - Lecture 4: Designing for two-stage samples
  - Lecture 5: Dealing with the real world – unequal sized clusters
  - Lecture 6: Subsampling
- Unit 4: Being more efficient
- Unit 5: Simplifying sampling
- Unit 6: Some extensions & applications



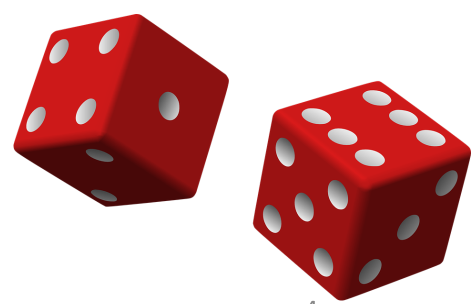
Unit 3

- 1 Simple complex
- 2 deff & roh
- 3 2-stage sampling
- 4 Designing 2-stage samples
- 5 Unequal sized clusters
- **6 Subsampling**

- Unit 1: Sampling as a research tool
- Unit 2: Mere randomization
- Unit 3: Saving money
  - Lecture 1: Simple complex sampling – choosing entire clusters
  - Lecture 2: Design effects & intraclass correlation
  - Lecture 3: Two-stage sampling
  - Lecture 4: Designing for two-stage samples
  - Lecture 5: Dealing with the real world – unequal sized clusters
  - **Lecture 6: Subsampling**
- Unit 4: Being more efficient
- Unit 5: Simplifying sampling
- Unit 6: Some extensions & applications



- Cost model
- Variance model
- Optimum subsample size
- Optimum number of clusters



4

- Cost model
- Variance model
- Optimum subsample size
- Optimum number of clusters

- Projecting standard errors and confidence intervals for cluster sampling depends on *b* and *deff*
- Estimating sample size for cluster sample sizes depends on *b* and *deff*
- That is, knowing *b* and *roh* leads to a projected *deff* & sample size *n*

$\boxed{n}$

- We know that as *b* goes up or down *deff* goes up or down
- And *var*(*p*) follows
- But we also have seen that as *b* goes up or down *a* goes down or up
- And as *a* goes down or up the cost of the data collection goes down or up
- There is a cost-error trade-off in cluster sample design

- Cost model
- Variance model
- Optimum subsample size
- Optimum number of clusters

- **Can we choose any set of *b* and *a*, as long as we don't exceed budget?**
- **Or is there a choice, an optimum choice for *a* and *b* that gives us the best (minimum sampling variance) among all possible choices for the given budget?**

optimum.

- **Cost model**
- Variance model
- Optimum subsample size
- Optimum number of clusters

- There is an "optimum" choice for *a* and *b*
- It can be obtained by minimizing the sampling variance for fixed cost (or vice versa)
- Cost model for two stage sampling:

$$C - C_0 = a\, c_a + a(b\, c_b)$$

- $C - C_0$ is the budget available, after overhead costs are removed

- c_a is the cost per cluster
- c_a is dominated by travel and preparation costs

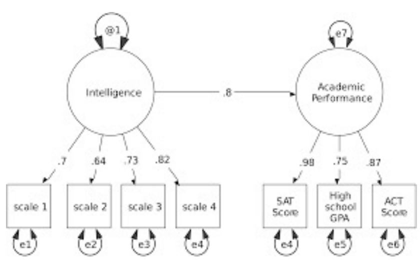
- c_b is the cost per observation within a cluster
- c_b is dominated by interviewing costs

- Cost model
- **Variance model**
- Optimum subsample size
- Optimum number of clusters

- **There is corresponding "sampling variance" model for two stage sampling:**

$$\text{var}(p) = \frac{(1-f)\,p(1-p)}{ab-1}\left[1+(b-1)roh\right]$$

- **As *a* goes up or down, the sampling variance goes up or down**
- **The relationship between *b* and sampling variance is more complicated …**



- Cost model
- Variance model
- Optimum subsample size
- Optimum number of clusters

- **The optimum subsample size for fixed cost $C - C_0$ can be found by a calculus or algebraic approach**
- **Finding b that minimizes the sampling variance**
- **The optimum b is**

$$b_{opt} = \sqrt{\frac{c_a}{c_b} \cdot \frac{1 - roh}{roh}}$$

- Cost model
- Variance model
- **Optimum subsample size**
- Optimum number of clusters

- **The optimum subsample size for fixed cost C - C_0 can be found by a calculus or algebraic approach**
- **Finding b that minimizes the sampling variance**

- **The optimum b is**

$$b_{opt} = \sqrt{\frac{c_a}{c_b} \cdot \frac{1-roh}{roh}}$$

- **As c_a increases, b increases**
- **As c_b increases, b decreases**
- **As *roh* increases, b decrases**

optimum.

- Cost model
- Variance model
- **Optimum subsample size**
- Optimum number of clusters

- **For example, if *roh* = 0.01, then** $\frac{1-roh}{roh}=\frac{1-0.01}{0.01}=\frac{0.99}{0.01}=99$
- **But if *roh* = 0.05, then** $\frac{1-roh}{roh}=\frac{1-0.05}{0.05}=\frac{0.95}{0.05}=19$
- **More homogeneity within, take fewer observations within …**

optimum.

- Cost model
- Variance model
- Optimum subsample size
- **Optimum number of clusters**

- **What about *a*?**

- **Consider the cost model again:**

$$C - C_o = ac_a + (ab_{opt})c_b$$

- **Solve for *a*:**

$$a = \frac{C - C_o}{c_a + b_{opt} c_b}$$

- Cost model
- Variance model
- Optimum subsample size
- **Optimum number of clusters**

- **For example, from a survey I once worked on, c_a = \$65.40 and c_b = \$25**
- **If *roh* = 0.05 (for a single variable, or on average),**

$$b_{opt} = \sqrt{\frac{65.40}{25} \cdot \frac{1-0.05}{0.05}} = 7.05$$

- **And if we had $C - C_0$ = \$10,000, then**

$$a = \frac{C - C_o}{c_a + b_{opt} c_b} = \frac{\$10,000}{\$65.40 + 7.05 \times \$25} = 41.38 \approx 41$$

- **We might in this case increase *b* to obtain an integer value for *a* that meets the budget exactly**

Unit 4

- 1 Stratifica-tion
- 2 Sampling variance
- 3 Propor-tionate allocation
- 4 Dispropor-tionate allocations
- 5 Comapring strata
- 6 Number of strata

- Unit 1: Sampling as a research tool
- Unit 2: Mere randomization
- Unit 3: Saving money
  - Lecture 1: Simple complex sampling – choosing entire clusters
  - Lecture 2: Design effects & intraclass correlation
  - Lecture 3: Two-stage sampling
  - Lecture 4: Designing for two-stage samples
  - Lecture 5: Dealing with the real world – unequal sized clusters
  - Lecture 6: Subsampling
- **Unit 4: Being more efficient**
- Unit 5: Simplifying sampling
- Unit 6: Some extensions & applications