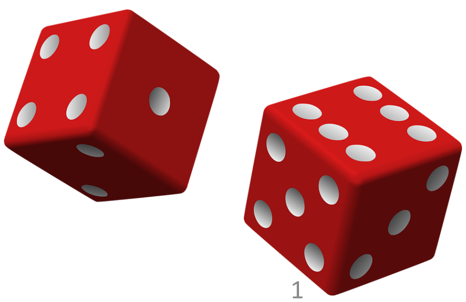
Sampling People, Records, & Networks

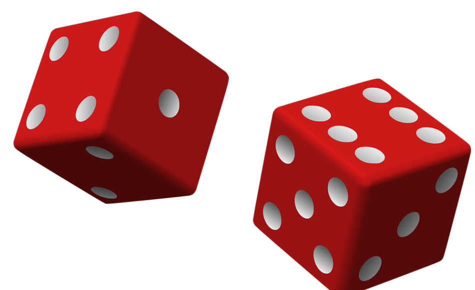
Jim Lepkowski, PhD
Professor & Research Professor *Emeritus*
Institute for Social Research, University of Michigan
Research Professor,
Joint Program in Survey Methodology, University of Maryland



Unit 3

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

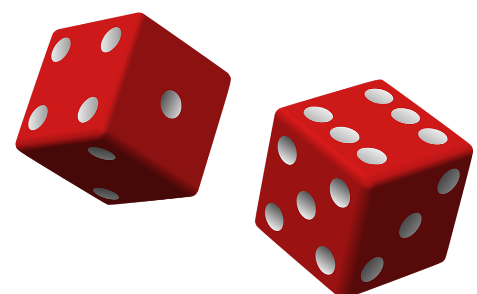
- Unit 1: Sampling as a research tool
- Unit 2: Mere randomization
- Unit 3: Saving money
    - Lecture 1: Simple complex sampling – choosing entire clusters
    - Lecture 2: Design effects & intraclass correlation
    - Lecture 3: Two-stage sampling
    - Lecture 4: Designing for two-stage samples
    - 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

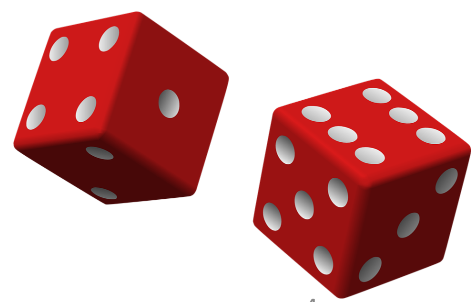
- 1 Simple complex
- 2 deff & roh
- 3 2-stage sampling
- 4 Designing 2-stage samples
- **5 Unequal sized clusters**
- 6 Subsampling

- 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



- The problem
- Sampling schemes
- PPS
- Systematic PPS

- 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



 4

- **The problem**
- Sampling schemes
- PPS
- Systematic PPS

- **Naturally occurring clusters tend to be unequal in size**
- **Fixed sampling rates and unequal sized clusters result in variation in sample size**

- **The problem**
- Sampling schemes
- PPS
- Systematic PPS

Hospital	B_a	Hospital	B_a
1	420	7	60
2	180	8	60
3	120	9	720
4	600	10	1860
5	240	11	1140
6	360	12	240

- **The problem**
- Sampling schemes
- PPS
- Systematic PPS

- **An *epsem* sample of n = 100 employees is desired from the N = 6,000**
  - **Select a = 2 hospitals**
  - $f = 100/6000 = 1/60$
- **First select SRS a = 2 (a rate of 1/6)**
  - **And then choose employees at the rate 1/10 within the selected hospitals**

$$f = (1/6)\cdot(1/10) = 1/60$$

- The problem
- Sampling schemes
- PPS
- Systematic PPS

- **Suppose hospitals 2 and 6 are chosen**
  - **Subsampling at the rate of 1/10 yields sample size**

    $$n = (180 + 360)/10 = 18 + 36 = 54$$

  - **If hospitals 2 and 10 were chosen, though,**

    $$n = (180 + 1860)/10 = 18 + 186 = 204$$

- **Subsample size varies**
- **Sample administration becomes difficult**

- **The problem**
- Sampling schemes
- PPS
- Systematic PPS

- **Variation in the overall sample size is undesirable**
- **Since *n* is a random variable, $\bar{y} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} y_i$ no longer applies**
- **We need to use a ratio estimator**

$$r = \frac{\sum_{\alpha=1}^{a} y_\alpha}{\sum_{\alpha=1}^{a} x_\alpha} = \frac{y}{x}$$

- The problem
- Sampling schemes
- PPS
- Systematic PPS

- **Seeking to control** $x = \sum_{\alpha=1}^{n} x_\alpha$
- **Controlled sample size provides administrative convenience in fieldwork**
- **Also provides greater statistical efficiency of estimators**
- **Several methods**
  - **Select exactly *b* elements per cluster**
  - **Probability proportionate to size (PPS)**

- Suppose a = 2 and b = 50 employees per selected hospital are chosen
  - Sample size is n = 100, and does not vary by which hospitals are chosen
- This design will on average across all possible samples over-represent employees in small hospitals
  - The probability of selection of small hospital employees is higher

- The problem
- Sampling schemes
- PPS
- Systematic PPS

- For example, for hospital 2,

$$f = (1/6)(50/180) = 1/21.6$$

- While for hospital 10,

$$f = (1/6)(50/1860) = 1/223.2$$

- This variation in rates can be remedied through weighting
  - Return to weighting in a later section

- The problem
- Sampling schemes
- PPS
- Systematic PPS

- Require a method that is
  - *Epsem*
  - Achieves equal sized subsamples in clusters
- Again, consider a = 2 and b = 50
- In order to achieve *epsem*, the following must be the "selection equation":

$$f = \frac{1}{60} = P\{\alpha\} \cdot \frac{50}{B_\alpha}$$

- **For example, if hospital 4 is chosen, then**

$$f = \frac{1}{60} = P\{\alpha\} \cdot \frac{50}{600} = P\{\alpha\} \cdot \frac{1}{12}$$

- **In order to make this *epsem*, we need**

$$\frac{1}{60} = P\{\alpha\} \cdot \frac{50}{B_\alpha}$$

$$P\{\alpha\} = \frac{1}{60} \cdot \frac{B_\alpha}{50} = \frac{B_\alpha}{3000}$$

- The problem
- Sampling schemes
- **PPS**
- Systematic PPS

- **Re-expressing,**

$$P\{\alpha\} = \frac{2 \cdot B_\alpha}{6000} = \frac{2 \cdot B_\alpha}{\sum_\alpha B_\alpha}$$

- **In general, this becomes, across two stages,**

$$f = P\{\alpha \ and \ \beta\} = \frac{a \cdot B_\alpha}{\sum_a B_\alpha} \cdot \frac{b}{B_\alpha} = \frac{a \cdot b}{\sum_a B_\alpha} = \frac{n}{N}$$

- The problem
- Sampling schemes
- **PPS**
- Systematic PPS

Hospital	B_α	Cum. B_α
1	420	420
2	180	600
3	120	720
4	600	1320
5	240	1560
6	360	1920
7	60	1980
8	60	2040
9	720	2760
10	1860	4620
11	1140	5760
12	240	6000

- The problem
- Sampling schemes
- **PPS**
- Systematic PPS

- Select RN’s from 1 to 6000, say …
  - RN = 702
  - RN = 1744
- Find the first hospital with cumulative sum greater than or equal to the first RN
- Find the next hospital with sum greater than the second RN
- These choose hospitals 3 and 7:

PPS

- The problem
- Sampling schemes
- **PPS**
- Systematic PPS

Hospital	B_α	Cum. B_α	
1	420	420	
2	180	600	
3	120	720	702
4	600	1320	
5	240	1560	
6	360	1920	1744
7	60	1980	
8	60	2040	
9	720	2760	
10	1860	4620	
11	1140	5760	
12	240	6000	

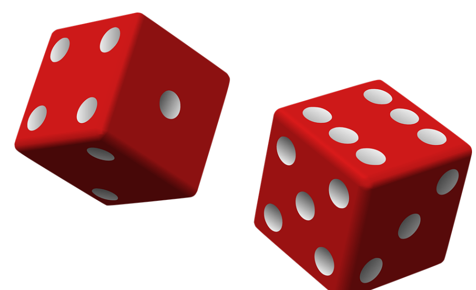
- The problem
- Sampling schemes
- PPS
- Systematic PPS

- Alternatively, select one RN from 1 to the interval 6000/2 = 3000
  - Say RN = 702
- Find the selected hospital, as above
- Add the interval to the RN to obtain 702 + 3000 = 3702
- Find the second hospital with this selection number, as above
- The RN 702 yields hospitals 3 and 10

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



20