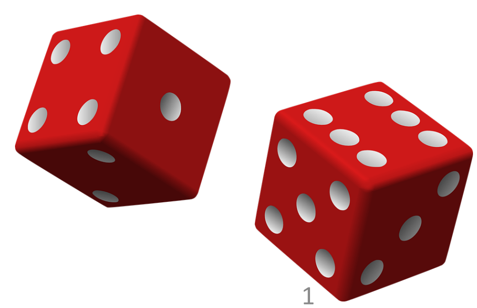
# Sampling People, Records, & Networks

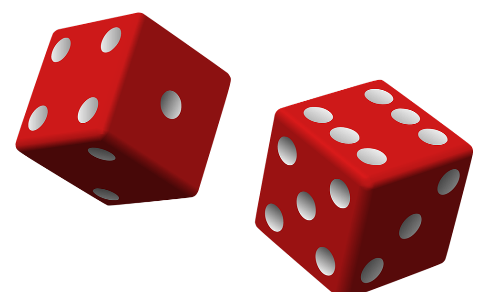
**Jim Lepkowski, PhD**
**Professor & Research Professor *Emeritus***
**Institute for Social Research, University of Michigan**
**Research Professor,**
**Joint Program in Survey Methodology, University of Maryland**



# 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

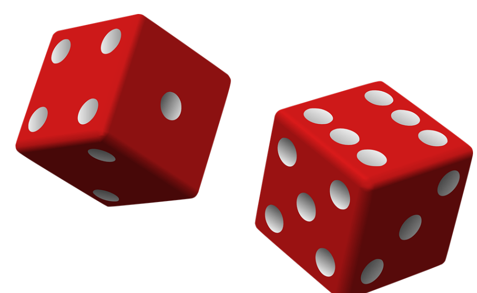



2

# Unit 3

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

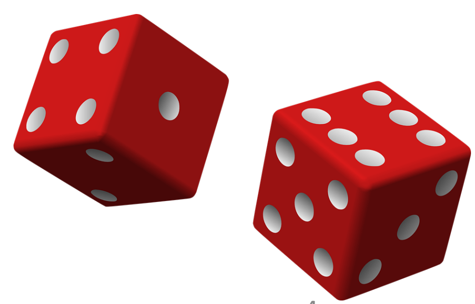
- Unit 1: Sampling as a research tool
- Unit 2: Mere randomization
- Unit 3: Saving money
  - Lecture 1: Simple complex sampling – choosing entire clusters
  - **Lecture 2: Design effects & intraclass correlation**
  - Lecture 3: Two-stage sampling
  - Lecture 4: Designing for two-stage samples
  - Lecture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




3

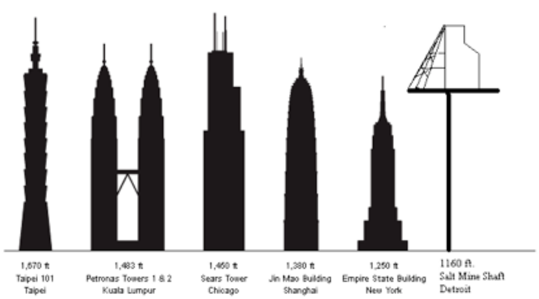
- deff
- roh
- Calculation

- 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



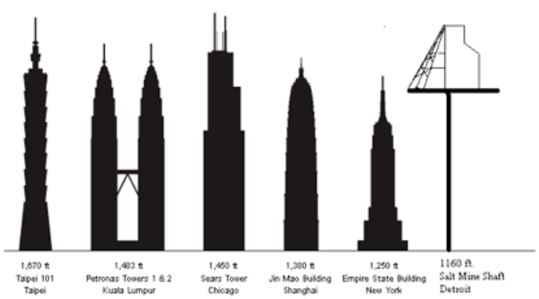
- **deff**
- roh
- Calculation

- **A question is how did the cluster sample compare to a simple random sample?**



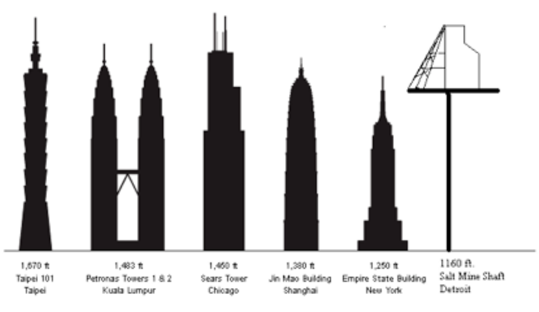
- **deff**
- roh
- Calculation

- A question is how did the cluster sample compare to a simple random sample?
- Need to establish grounds for comparison
  - Compare precision since both designs are unbiased, and yield the same mean on average
  - On what basis should the precision be compared?
  - Usually equal sample size
  - And a comparison of sampling variances



- **deff**
- roh
- Calculation

- **A question is how did the cluster sample compare to a simple random sample?**
- **Need to establish grounds for comparison**
  - **Compare precision since both designs are unbiased, and yield the same mean on average**
  - **On what basis should the precision be compared?**
  - **Usually equal sample size**
  - **And a comparison of sampling variances**
- **If the sample had instead been an SRS of *n* = 240 children from all schools, then**




$$p = 160/240$$

$$\operatorname{var}_{SRS}(p) = (1-f)\frac{p(1-p)}{n-1} = 0.0009112$$

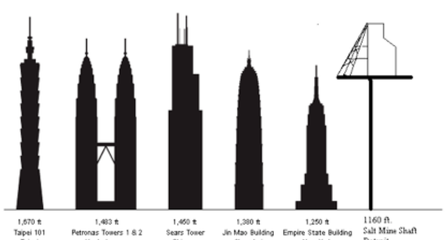
- **deff**
- roh
- Calculation

- **Compared to cluster sampling, the estimated variance of *p* is considerably smaller for SRS**
- **A ratio quantifies the comparison:**

$$deff(p) = \frac{\operatorname{var}(p)}{\operatorname{var}_{SRS}(p)}$$

- **By definition, the numerator sampling variance must have the same sample size as the denominator**
- **For the illustration,**

$$deff(p) = \frac{\operatorname{var}(p)}{\operatorname{var}_{SRS}(p)} = \frac{0.002760}{0.0009112} = 3.029$$

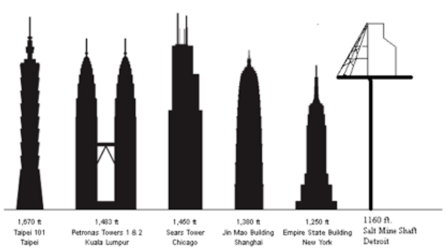


- **deff**
- roh
- Calculation


- **The design effect may be used in several ways**
- **One is to recognize that is says the following:**

$$\mathrm{var}\left(p\right)=deff\left(p\right)\times \mathrm{var}_{SRS}\left(p\right)$$

- **In other words, the cluster sampling variance is the SRS sampling variance, adjusted for the effect of clustering**
- **This expression can be used to help design new surveys – to be discussed in the next lecture**



- **deff**
- roh
- Calculation

- The design effect is directly a function of differences between clusters compared to differences among elements
- If deff > 1, then clusters are more variable than elements
- But why?

- deff
- roh
- Calculation

- The design effect is directly a function of differences between clusters compared to differences among elements
- If deff > 1, then clusters are more variable than elements
- But why?
- Heterogeneity between implies homogeneity within:

  The more different clusters are from one another … the more similar are elements within clusters to one another

- **deff**
- roh
- Calculation

- **Empirical results have revealed that deff depends on homogeneity within and the size of the clusters, say *b***
- **The homogeneity is measured by the intra-cluster correlation *roh***

- **deff**
- roh
- Calculation

- **Empirical results have revealed that deff depends on homogeneity within and the size of the clusters, say *b***
- **The homogeneity is measured by the intra-cluster correlation *roh***
- ***rate of homogeneity***
- **The design effect is given by**

$$deff(p) = 1 + (b-1)roh$$

- deff
- **roh**
- Calculation

- **The intra-cluster correlation can be estimated from the design effect:**

$$roh = \frac{deff\left(p\right) - 1}{b - 1}$$

$$= \frac{3.029 - 1}{24 - 1}$$

$$= 0.088$$

- deff
- roh
- Calculation

- *roh* is a property of the clusters and the variable under study
  - The design effect is then also going to differ across variables
- *roh* is substantive, not statistical
- *roh* is nearly always positive
  - Elements in a cluster tend to resemble one another
- Source of *roh*
  - Environment
  - Self-selection
  - Interaction

- deff
- roh
- Calculation

- Alternatively, the actual sample size is $n$ = 240 in the cluster sample
- But an SRS that is equally precise would only have to have

$$n_{eff} = \frac{240}{3.029} = 79$$

- Effective sample size

- deff
- roh
- Calculation

- Consider alternative outcomes for our sample of *a* = 10 classrooms
  - Homogeneity with, heterogeneity between

$$\frac{0}{24},\frac{0}{24},\frac{0}{24},\frac{16}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24}$$

- deff
- roh
- Calculation

- **Consider alternative outcomes for our sample of *a* = 10 classrooms**
  - **Homogeneity with, heterogeneity between**

$$\frac{0}{24},\frac{0}{24},\frac{0}{24},\frac{16}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24}$$

$$s_a^2 = 0.2222 \quad \mathrm{var}\left(p\right) = 0.02178$$

- deff
- roh
- Calculation

- **Consider alternative outcomes for our sample of *a* = 10 classrooms**
  - **Homogeneity with, heterogeneity between**

$$\frac{0}{24},\frac{0}{24},\frac{0}{24},\frac{16}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24}$$

$$s_a^2 = 0.2222 \quad \text{var}\left(p\right) = 0.02178$$

$$deff = 23.90$$

- deff
- roh
- Calculation

- **Consider alternative outcomes for our sample of *a* = 10 classrooms**
  - **Homogeneity with, heterogeneity between**

$$\frac{0}{24},\frac{0}{24},\frac{0}{24},\frac{16}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24}$$

$$s_a^2 = 0.2222 \quad \operatorname{var}(p) = 0.02178$$

$$deff = 23.90$$

$$n_{eff} = 240 / 23.9 = 10$$

- **Consider alternative outcomes for our sample of *a* = 10 classrooms**
  - **Homogeneity with, heterogeneity between**

$$\frac{0}{24},\frac{0}{24},\frac{0}{24},\frac{16}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24}$$

$$s_a^2 = 0.2222 \quad \mathrm{var}\left(p\right) = 0.02178$$

$$deff = 23.90$$

$$n_{eff} = 240 / 23.9 = 10$$

$$roh = \frac{23.90 - 1}{24 - 1} = 0.996$$

- deff
- **roh**
- Calculation

- **Heterogeneity within, homogeneity between**

$$\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24}$$

- deff
- **roh**
- Calculation

- **Heterogeneity within, homogeneity between**

$$\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24}$$

$$s_a^2 = 0.0 \quad \mathrm{var}\left(p\right) = 0.0$$

$$deff = 0$$

$$n_{eff} = 240 / 0$$

- deff
- **roh**
- Calculation

- **Heterogeneity within, homogeneity between**

$$\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24}$$

$$s_a^2 = 0.0 \quad \mathrm{var}\left(p\right) = 0.0$$

$$deff = 0$$

$$n_{eff} = 240/0$$

$$roh = \frac{0-1}{24-1} = -0.043$$

- **Consider an equal probability (*epsem*) sample of *n* = 2,400 obtained from a one-stage sample of *a* = 60 equal-sized clusters each of size *b* = 40 selected by SRS**
- **In a journal article describing survey results, for a key proportion, *p* = 0.40** $\text{var}\left(p\right) = 0.00021795$

**How would we estimate *deff* and *roh*?**

- deff
- roh
- **Calculation**

1. **Compute the simple random sampling variance**

$$\mathrm{var}_{SRS}\left(p\right)=\frac{p\left(1-p\right)}{n-1}$$

**(Ignore the *fpc* – that is, or assume it is 1)**

2. **Compute the design effect**

$$deff\left(p\right)=\frac{\mathrm{var}_{SRS}\left(p\right)}{\mathrm{var}\left(p\right)}=\frac{0.00021795}{\frac{p\left(1-p\right)}{n-1}}$$

3. **Compute the intra-cluster homogeneity *roh***

$$roh=\frac{deff\left(p\right)-1}{b-1}=\frac{deff\left(p\right)-1}{40-1}=$$

- deff
- roh
- **Calculation**

## The SRS variance is

$$\operatorname{var}_{SRS}\left(p\right)=\frac{p\left(1-p\right)}{n}=\frac{0.4\times 0.6}{2400}=0.0001$$

- deff
- roh
- **Calculation**

## The SRS variance is

$$\mathrm{var}_{SRS}\left(p\right)=\frac{p\left(1-p\right)}{n}=\frac{0.4\times 0.6}{2400}=0.0001$$

## Thus, the design effect is

$$deff\left(p\right)=\frac{\mathrm{var}\left(p\right)}{\mathrm{var}_{SRS}\left(p\right)}=\frac{0.00021795}{0.0001}=2.1795$$

- deff
- roh
- **Calculation**

## The SRS variance is

$$\text{var}_{SRS}(p) = \frac{p(1-p)}{n} = \frac{0.4 \times 0.6}{2400} = 0.0001$$

## Thus, the design effect is

$$deff(p) = \frac{\text{var}(p)}{\text{var}_{SRS}(p)} = \frac{0.00021795}{0.0001} = 2.1795$$

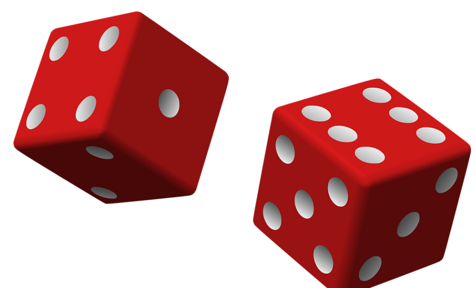
## And an estimate of intra-class correlation is

$$roh = \frac{deff(p)\text{ - }1}{b\text{ - }1} = \frac{2.1795\text{ - }1}{40\text{ - }1} = 0.03024$$

# 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



 30