Lecture 22: Network scale-up method to study groups most at-risk for HIV

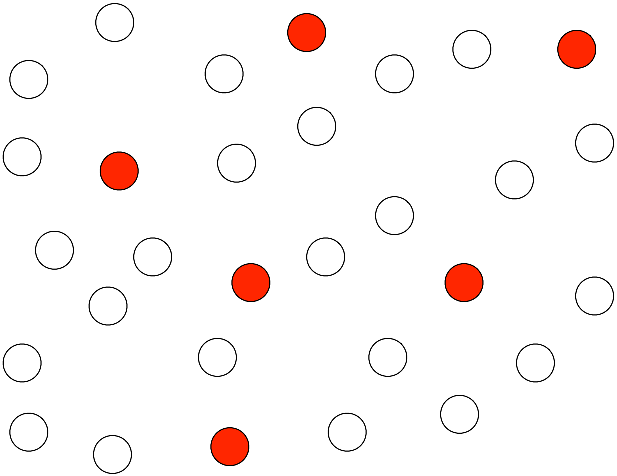
Matthew J. Salganik

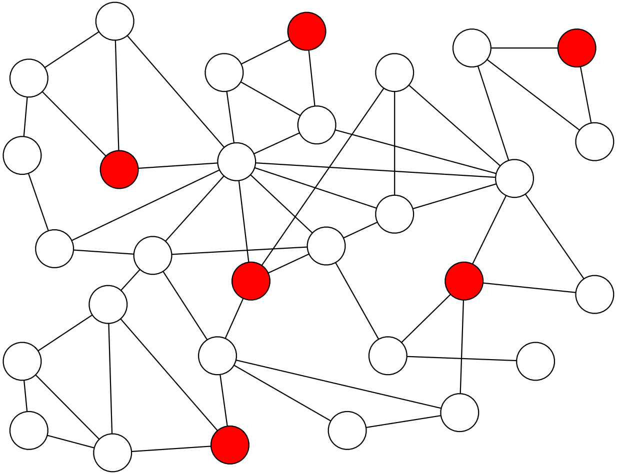
Sociology 204: Social Networks
Princeton University

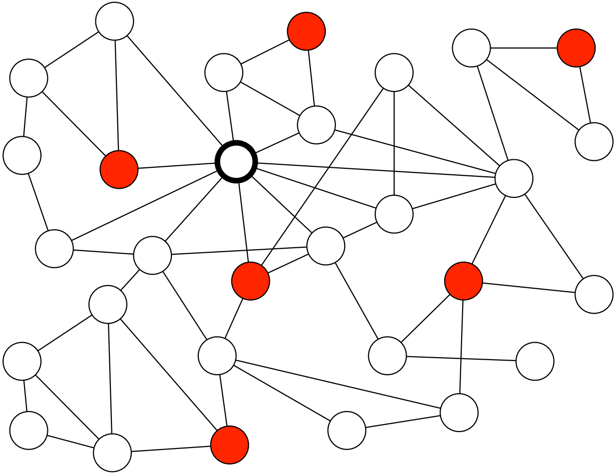
1/2 Network scale-up method, methods

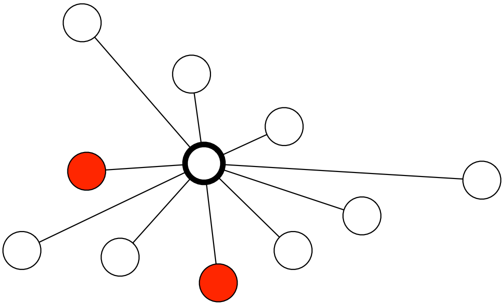
Network scale-up method

Basic insight from Bernard et al. (1989)



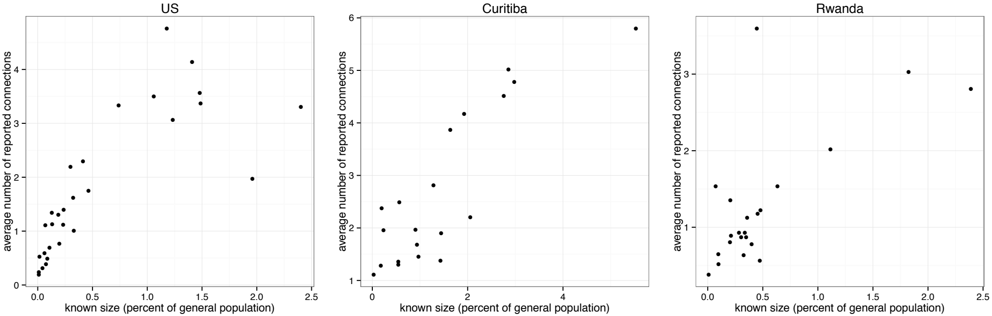






$$\hat{N}_H = \frac{2}{10} \times 30 = 6$$

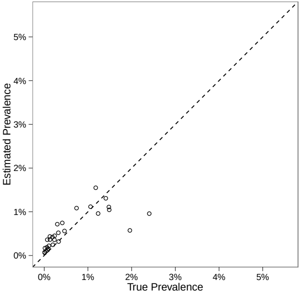
- Requires a random sample from the entire population
- Respondents are asked:
  - How many people do you know who are drug injectors?
  - How many women do you know that have given birth in the last 12 months?
  - How many people do you know who are middle school teachers?
  - ...
  - How many people do you know named Michael?
- "Know" typically defined: you know them and they know you and have you been in contact with them over the past two years



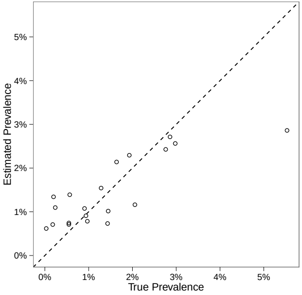
On average, these answers are not crazy.

Other size estimation methods are problematic, and scale-up method has many nice properties:

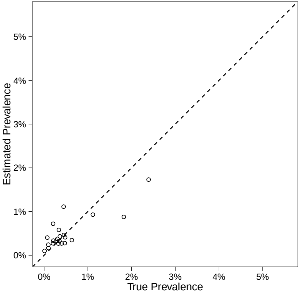
- Requires a random sample of the general population, not specific contact with the hard-to-reach population
- Can be added as a module (5-10 minutes) in any existing survey
- Can estimate many target populations in a single survey
- Can be applied at the city-level, sub-national-level, or national-level
- Statistical methods are improvable
- Partially self-validating because it uses groups of known size



(a) United States



(b) Curitiba



(c) Rwanda

But, basic scale-up also has problems. I will focus on insights about basic scale-up that we discovered from developing the generalized scale-up.

Personal background

How did I end up working on this research?

How did I end up working on this research?

- Working on sampling at the Census Bureau

How did I end up working on this research?

- Working on sampling at the Census Bureau
- When I began grad school I knew about sampling and was interested in networks and Doug Heckathorn was working on respondent-driven sampling

How did I end up working on this research?

- Working on sampling at the Census Bureau
- When I began grad school I knew about sampling and was interested in networks and Doug Heckathorn was working on respondent-driven sampling
- Respondent-driven sampling lead to the network scale-up method

Modeling
counting with multiplicity

$\leftrightarrow$

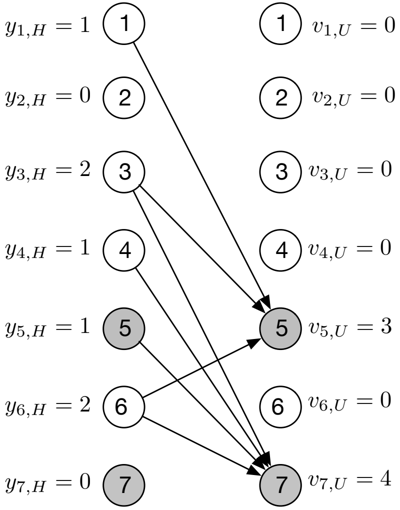
Empirical
Rwanda (this lecture)
Brazil (next lecture)

If $\underbrace{y_{i,k} \sim Bin(d_i, N_k/N)}_{\text{basic scale-up model}}$, then maximum likelihood estimator is

$$\hat{N}_H = \frac{\sum_i y_{i,H}}{\sum_i \hat{d}_i} \times N$$

- $\hat{N}_H$: number of people in the hidden population
- $y_{i,H}$: number of people in hidden population known by person i
- $\hat{d}_i$: estimated number of people known by person i
- N: number of people in the population

See Killworth et al., (1998)


$y_{1,H} = 1$
1
$y_{2,H} = 0$
2
$y_{3,H} = 2$
3
$y_{4,H} = 1$
4
$y_{5,H} = 1$
5
$y_{6,H} = 2$
6
$y_{7,H} = 0$
7
1
$v_{1,U} = 0$
2
$v_{2,U} = 0$
3
$v_{3,U} = 0$
4
$v_{4,U} = 0$
5
$v_{5,U} = 3$
6
$v_{6,U} = 0$
7
$v_{7,U} = 4$

$$\text{total out-reports} = \text{total in-reports}$$

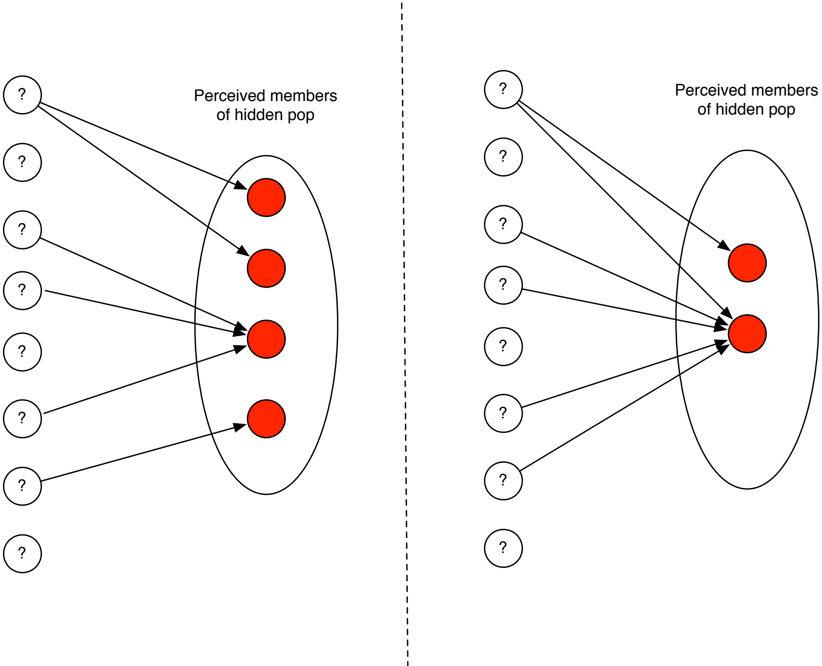
$$\text{total out-reports} = \text{total in-reports}$$

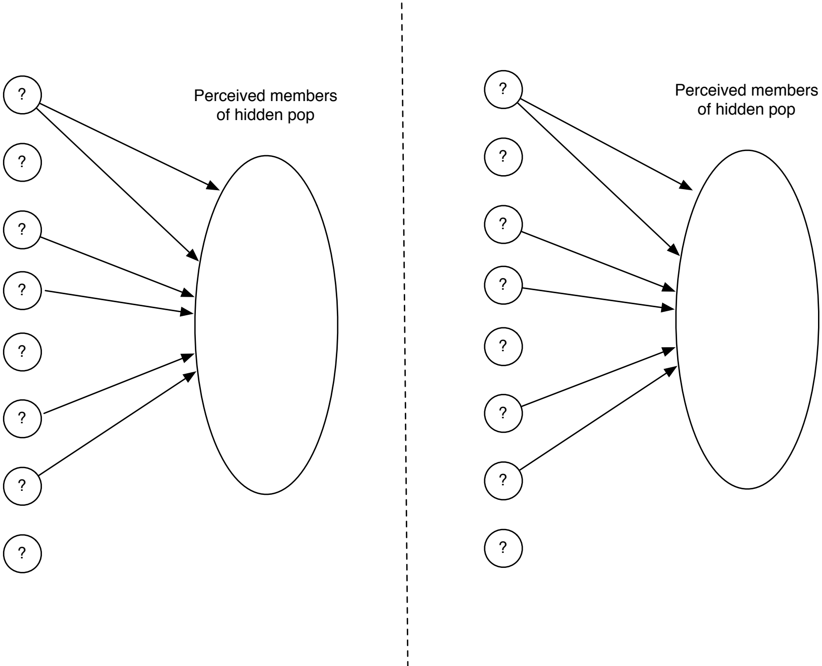
$$\begin{aligned}\text{total out-reports} = {} & \text{size of hidden pop} \times \\ & \text{in-reports per member of hidden pop}\end{aligned}$$

$$\text{total out-reports} = \text{total in-reports}$$

$$\text{total out-reports} = \text{size of hidden pop} \times \text{in-reports per member of hidden pop}$$

$$\text{size of hidden pop} = \frac{\text{total out-reports}}{\text{in-reports per member of hidden pop}}$$


Perceived members of hidden pop
?
?
?
?
?
?
?
?
Perceived members of hidden pop
?
?
?
?
?
?
?
?



Perceived members
of hidden pop
?
?
?
?
?
?
?
?
Perceived members
of hidden pop
?
?
?
?
?
?
?
?

$$\underbrace{N_H}_{\text{size of hidden pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,H}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{i \in U} v_{i,F} / N_H\right)}_{\text{in-reports per member of hidden pop}}}$$

$$\underbrace{N_H}_{\text{size of hidden pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,H}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{i \in U} v_{i,F} / N_H\right)}_{\text{in-reports per member of hidden pop}}}$$

If there are no false positives,

$$\underbrace{N_H}_{\text{size of hidden pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,H}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{i \in H} v_{i,F} / N_H\right)}_{\text{avg visible degree of hidden pop}}}$$

Generalized scale-up identity

$$\underbrace{N_H}_{\text{size of hidden pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,H}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{\color{blue}{i \in H}} v_{i,F} / N_H \right)}_{\text{avg visible degree of hidden pop}}}$$

Basic scale-up estimator

$$\hat{N}_H = \frac{\sum_{i \in s_F} y_{i,H}}{\sum_{i \in s_F} \hat{d}_i} \times N$$

Generalized scale-up identity

$$\underbrace{N_H}_{\text{size of hidden pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,H}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{\color{blue}{i \in H}} v_{i,F} / N_H \right)}_{\text{avg visible degree of hidden pop}}}$$

Basic scale-up estimator

$$\underbrace{\hat{N}_H}_{\text{est size of hidden pop}} = \frac{\overbrace{\sum_{i \in s_F} y_{i,H}}^{\text{total out-reports}}}{\underbrace{\sum_{i \in s_F} \hat{d}_{i,U} / N}_{\text{avg degree of pop}}}$$

Counting with multiplicity approach:

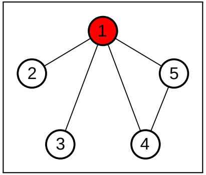
- no assumptions about the underlying social network
- extends naturally to incomplete social awareness
- extends naturally to incomplete frames
- extends naturally to complex sample designs

$$N_H = \underbrace{\left(\frac{y_{F,H}}{\bar{d}_{U,F}}\right)}_{\substack{\text{basic}\\ \text{scale-up}}} \times \underbrace{\underbrace{\frac{1}{\bar{d}_{F,F}/\bar{d}_{U,F}}}_{\substack{\text{frame ratio}\\ \phi_F}} \times \underbrace{\frac{1}{\bar{d}_{H,F}/\bar{d}_{F,F}}}_{\substack{\text{degree ratio}\\ \delta_F}} \times \underbrace{\frac{1}{\bar{v}_{H,F}/\bar{d}_{H,F}}}_{\substack{\text{true positive rate}\\ \tau_F}}}_{\text{adjustment factors}} = \underbrace{\left(\frac{y_{F,H}}{\bar{v}_{H,F}}\right)}_{\substack{\text{generalized}\\ \text{scale-up}}}. \tag{15}$$

Frame ratio: Less a focus for us. As a first approximation, sampling frame is adults you don't want to include kids in any of the reports

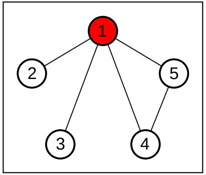
Degree ratio: If the hidden population has smaller network sizes than the general population, the size of the hidden population will be underestimated. Likewise, if the hidden population has larger network sizes than the general population, the size of the hidden population will be overestimated.

Set of egos can be different from set of alters.

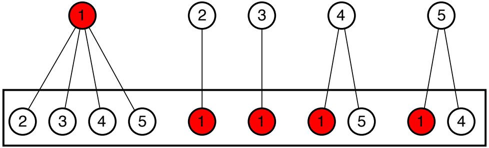


$p = 0.2$

Set of egos can be different from set of alters.

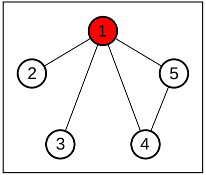


$p = 0.2$

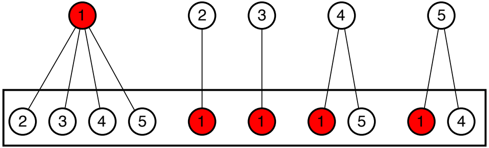


$p_{alter} = 0.4$

Set of egos can be different from set of alters.



$p = 0.2$



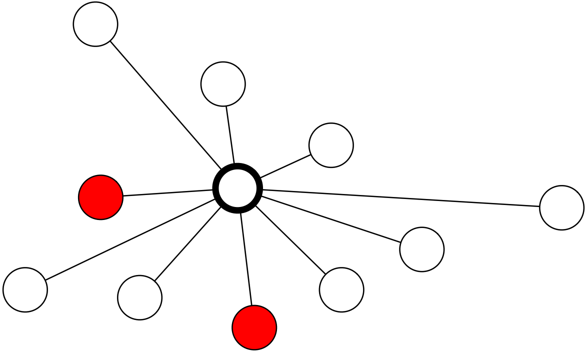
$p_{alter} = 0.4$

$$p_{alter} = p \times \frac{\text{avg. degree (hidden pop.)}}{\text{avg. degree (general pop.)}} = p\delta$$

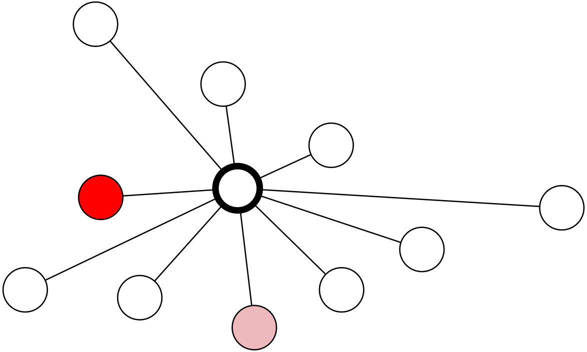
Estimates will be biased by a factor of δ_F ("degree ratio")

True positive rate: If people are connected to people in the hidden population but not aware of it, the size of the hidden population will be underestimated

How might imperfect knowledge impact scale-up estimates?
Ego is not aware of everything about all of their alters.



How might imperfect knowledge impact scale-up estimates?
Ego is not aware of everything about all of their alters.



Estimates will be biased by a factor of τ_F ("true positive rate")

More about this in next lecture

$$N_H = \underbrace{\left(\frac{y_{F,H}}{\bar{d}_{U,F}}\right)}_{\substack{\text{basic}\\ \text{scale-up}}} \times \underbrace{\underbrace{\frac{1}{\bar{d}_{F,F}/\bar{d}_{U,F}}}_{\substack{\text{frame ratio}\\ \phi_F}} \times \underbrace{\frac{1}{\bar{d}_{H,F}/\bar{d}_{F,F}}}_{\substack{\text{degree ratio}\\ \delta_F}} \times \underbrace{\frac{1}{\bar{v}_{H,F}/\bar{d}_{H,F}}}_{\substack{\text{true positive rate}\\ \tau_F}}}_{\text{adjustment factors}} = \underbrace{\left(\frac{y_{F,H}}{\bar{v}_{H,F}}\right)}_{\substack{\text{generalized}\\ \text{scale-up}}}. \qquad (15)$$

From a talk I gave at a UNAIDS workshop

Generalized scale-up approach

- simple estimators (just addition, subtraction, multiplication, and division)
- handles incomplete social awareness
- no assumptions about the underlying social network
- handles incomplete frames
- handles complex sample designs

and could still be very wrong in practice!

Assumptions can be put into four broad categories

- sampling
- social network structure
- reporting
- survey construction

Results for non-sampling assumptions have this form:

Estimator	Imperfect assumptions	Effective estimand
$\widehat{\bar{d}}_{F,F}$ (Result B.3)	(i) $\widehat{N}_{\mathcal{A}} = c_1 \, N_{\mathcal{A}}$	$\frac{c_2 \, c_3}{c_1} \, \bar{d}_{F,F}$
	(ii) $\bar{d}_{\mathcal{A},F} = c_2 \, \bar{d}_{F,F}$	
	(iii) $y_{F,\mathcal{A}} = c_3 \, d_{F,\mathcal{A}}$	

Estimator	Imperfect assumptions	Effective estimand
$\widehat{\bar{d}}_{F,F}$ (Result B.3)	(i) $\widehat{N}_{\mathcal{A}} = c_1\ N_{\mathcal{A}}$ (ii) $\bar{d}_{\mathcal{A},F} = c_2\ \bar{d}_{F,F}$ (iii) $y_{F,\mathcal{A}} = c_3\ d_{F,\mathcal{A}}$	$\frac{c_2\ c_3}{c_1}\ \bar{d}_{F,F}$
$\widehat{\bar{d}}_{U,F}$ (Result B.4)	(i) $\widehat{N}_{\mathcal{A}} = c_1\ N_{\mathcal{A}}$ (ii) $\bar{d}_{\mathcal{A},F} = c_2\ \bar{d}_{U,F}$ (iii) $y_{F,\mathcal{A}} = c_3\ d_{F,\mathcal{A}}$	$\frac{c_2\ c_3}{c_1}\ \bar{d}_{U,F}$
$\widehat{\phi}_F$ (Result B.6)	(i) $\widehat{\bar{d}}_{F,F} \rightsquigarrow c_1\ \bar{d}_{F,F}$ (ii) $\widehat{\bar{d}}_{U,F} \rightsquigarrow c_2\ \bar{d}_{U,F}$	$\frac{c_1}{c_2}\ \phi_F$
$\widehat{\bar{v}}_{H,F}$ (Result C.2)	(i) $\widehat{N}_{\mathcal{A}\cap F} = c_1\ N_{\mathcal{A}\cap F}$ (ii) $\tilde{v}_{H,\mathcal{A}\cap F} = c_2\ v_{H,\mathcal{A}\cap F}$ (iii) $\frac{v_{H,\mathcal{A}\cap F}}{N_{\mathcal{A}\cap F}} = c_3\ \frac{v_{H,F}}{N_F}$	$\frac{c_3\ c_2}{c_1}\ \bar{v}_{H,F}$
$\widehat{\delta}_F$ (Result C.6)	(i) $\widehat{\bar{d}}_{H,F} \rightsquigarrow c_1\ \bar{d}_{H,F}$ (ii) $\widehat{\bar{d}}_{F,F} \rightsquigarrow c_2\ \bar{d}_{F,F}$	$\frac{c_1}{c_2}\ \delta_F$
$\widehat{\tau}_F$ (Result C.7)	(i) $\widehat{\bar{v}}_{H,F} \rightsquigarrow c_1\ \bar{v}_{H,F}$ (ii) $\widehat{\bar{d}}_{H,F} \rightsquigarrow c_2\ \bar{d}_{H,F}$	$\frac{c_1}{c_2}\ \tau_F$
$\widehat{N}_H$ (Result C.8)	(i) $\widehat{\bar{v}}_{H,F} \rightsquigarrow c_1\ \bar{v}_{H,F}$	$\frac{1}{c_1}\ N_H$
$\widehat{N}_H$ (Result C.10)	(i) $\widehat{\bar{d}}_{F,F} \rightsquigarrow c_1\ \bar{d}_{F,F}$ (ii) $\widehat{\delta}_F \rightsquigarrow c_2\ \delta_F$ (iii) $\widehat{\tau}_F \rightsquigarrow c_3\ \tau_F$	$\frac{1}{c_1\ c_2\ c_3}\ N_H$

Fails when:

- adjustment factors not measured
- adjustment factors measured poorly
- reporting not consistent with awareness
- variance estimation fails
- . . . .

Modeling		Empirical
counting with multiplicity	$\leftrightarrow$	Rwanda (this lecture)
		Brazil (next lecture)