Capture-Recapture Methods

sampling approach to estimate an unknown population size by using two or more samples from that population

take at least samples from the population and mark those with the desired trait

	Not in Sample 2	**In Sample 2**	
Not in Sample 1	n_{00}	n_{01}	
In Sample 1	n_{10}	n_{11}	N_1
		N_2	N

n_{00} = total that are not captured
N = total population
n_{00} and N are unknown

Assumptions

1) closed population — population doesn't change between first and second samples
2) no lost marks between samples — can always successfully match those captured twice
3) homogenous probability of capture — all members of the population have equal probability of being captured
4) source independence — probability of first and second capture are independent

Lincoln-Peterson Estimator

n_1 = number captured in sample 1
n_2 = number captured in sample 2
n_{11} = number captured n both samples

$$p(1+) = \frac{N_1}{N}$$

$$p(2+) = \frac{N_2}{N}$$

$$p(1 + and\ 2 +) = \frac{N_1}{N} \times \frac{N_2}{N} = \frac{n_{11}}{N}$$

$$\hat{N} = \frac{N_1 N_2}{n_{11}}$$

$$\hat{x} = \frac{(N_1 - n_{11})(N_2 - n_{11})}{n_{11}} = \frac{n_{10} \times n_{01}}{n_{11}}$$

Log Linear Models/Poisson Regression Models

outcome is a count and has a Poisson distribution

$$\ln(n) = \beta_0 + \beta_1 Source1 + \beta_2 Source2$$
$$n_{00} = e^{\beta_0}$$
$$n_{10} = e^{\beta_0+\beta_1}$$
$$n_{01} = e^{\beta_0+\beta_2}$$
$$n_{11} = e^{\beta_0+\beta_1+\beta_2}$$

β

Violation of Assumptions

<u>Assumption 1</u> – Closed population
may be invalid if the two captures occur far apart in time
surveillance data are often organized around convenient annual periods

<u>Assumption 2</u> – No lost marks between samples
large amounts of missed or incorrect matches between databases will bias estimates

<u>Assumption 3</u> – Homogeneity of capture probabilities
probability of being captured by a data sources depends in individual covariates, e.g., gender, insurance coverage
stratify and estimate populations within each stratum
model probability of capture conditional on observed covariates
accounting for heterogeneity may remove source dependence

<u>Assumption 4</u> – Source Independence
data sources are not independent of each other

- Positive Source dependence
  results in underestimation
  e.g., someone who is engaged in the system are more likely to show up in multiple datasets

$$p(1 + and\ 2 +) \geq P(1 +) \times\ P(2 +)$$
$$\frac{n_{12}}{N} \geq \frac{N_1}{N} \times \frac{N_2}{N}$$
$$N \geq \frac{N_1 N_2}{n_{12}}$$

- Negative Source dependence
  results in overestimation
  degree of mutual exclusivity between lists

$$p(1 + and\ 2 +) \leq P(1 +) \times\ P(2 +)$$
$$\frac{n_{12}}{N} \leq \frac{N_1}{N} \times \frac{N_2}{N}$$
$$N \leq \frac{N_1 N_2}{n_{12}}$$

Interaction Terms

when there are more than two data sources, log linear model can include an interaction term to represent model dependence between sources

cannot model dependence between all sources