

# The Integrated Census in Israel

Using Sample Surveys to Estimate  
Coverage Errors in Administrative Data

**Hagit Glickman, Ronit Nirel, Dan Ben-Hur**

**Central Bureau of Statistics, Israel**

May 2005

## Introduction:

---

- The basic idea of the Integrated Census is:
  - ✓ Replace the traditional nationwide field enumeration with an administrative enumeration as the basis for population estimates
  - ✓ Correct and augment the administrative data using information obtained from sample surveys.

# Introduction:

---

- Expected gains:
  - ✓ Improved quality and timeliness of estimates
  - ✓ Reduction in response burden
  - ✓ Increase in census frequency
  - ✓ Reduced cost

# The Population Register:

---

- The main administrative source is the national Population Register (PR).

Information provided by the PR includes:

- ✓ A unique ID number and name
- ✓ An address
- ✓ Basic demographic information: age, sex, place of birth, date of immigration, race/ethnicity, marital status, religion, and kinship relation

## The Population Register:

---

- Coverage errors of the PR include:
  - ✓ Local undercoverage and overcoverage due to outdated addresses
  - ✓ National overcoverage of emigrants still listed in the PR.
  - ✓ National undercoverage of people living in Israel without an ID number, legally or illegally.

The extent of coverage errors is differential across geographical areas and demographic characteristics.

## Estimation objective:

---

- The Israeli administrative-statistical system divides the country into localities and statistical areas (census tracts) within localities. A statistical area comprises on average 4000 residents. Localities having less than 10,000 residents are regarded as a single statistical area.
- The Integrated Census is designed to provide accurate population estimates for statistical areas.

## The coverage model :

---

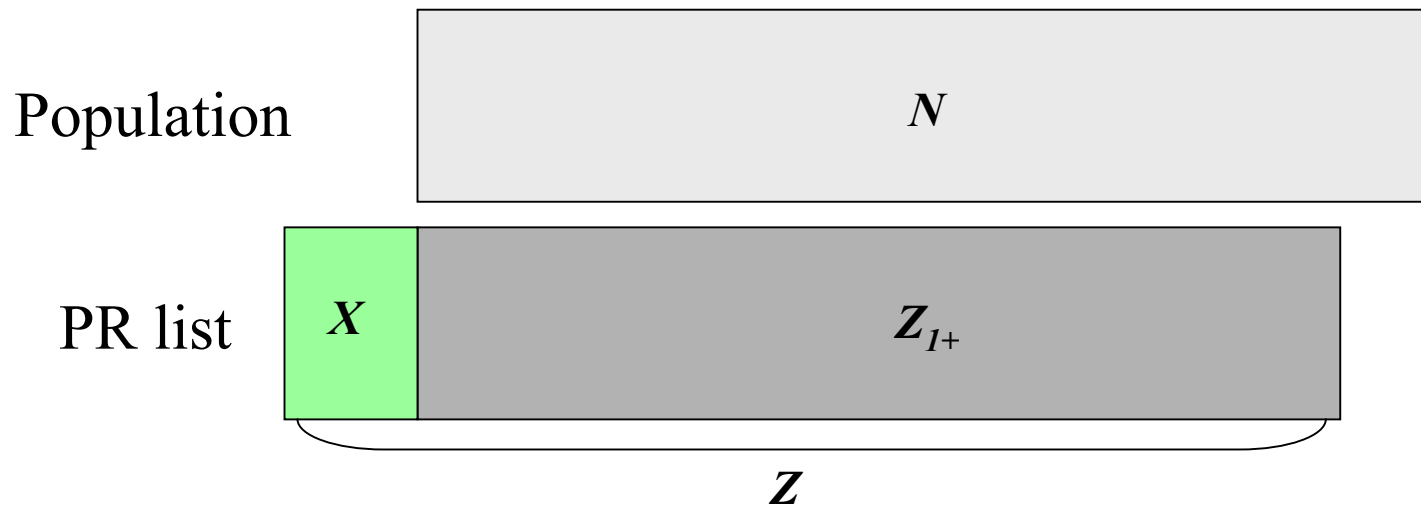
- Coverage errors are defined with respect to a statistical area. That is,

The PR undercoverage for a given statistical area is composed of all persons living in that area but listed elsewhere in the PR.

The PR overcoverage for a given statistical area is composed of all persons listed in the PR in that area but live elsewhere (either in Israel or abroad).

# The coverage model:

---



The coverage parameters are defined as

$$p_{1+} = EZ_{1+} / N \quad \lambda = EX / N$$

## The coverage model:

---

Note that  $p_{1+} + \lambda = EZ / N$ .

Since the coverage parameters,  $p_{1+}$  and  $\lambda$ , are unknown, we estimate them using sample surveys.

The estimate of the population size is

$$\hat{N} = \frac{Z}{\hat{p}_{1+} + \hat{\lambda}}$$

## Coverage samples:

---

- In order to estimate the coverage parameters, two sample surveys are designed:
  - ✓ Area-based sample to estimate the undercoverage parameter - **U sample**
  - ✓ Sample of people from the PR to estimate the overcoverage parameter - **O sample**

## Coverage samples:

---

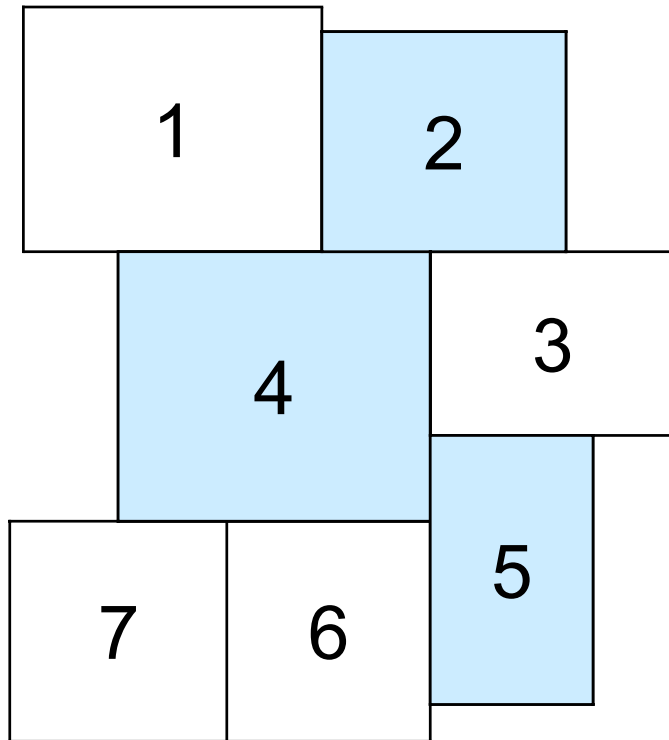
The two samples are obtained as follows:

- Statistical areas are divided into enumeration areas (EAs), each comprising around 50 households.
- PR addresses are geocoded and clustered by EAs.
- The same EAs are selected for both samples.

# Coverage sample:

---

Area



**U sample**

PR



**O sample**

## Coverage samples - U sample

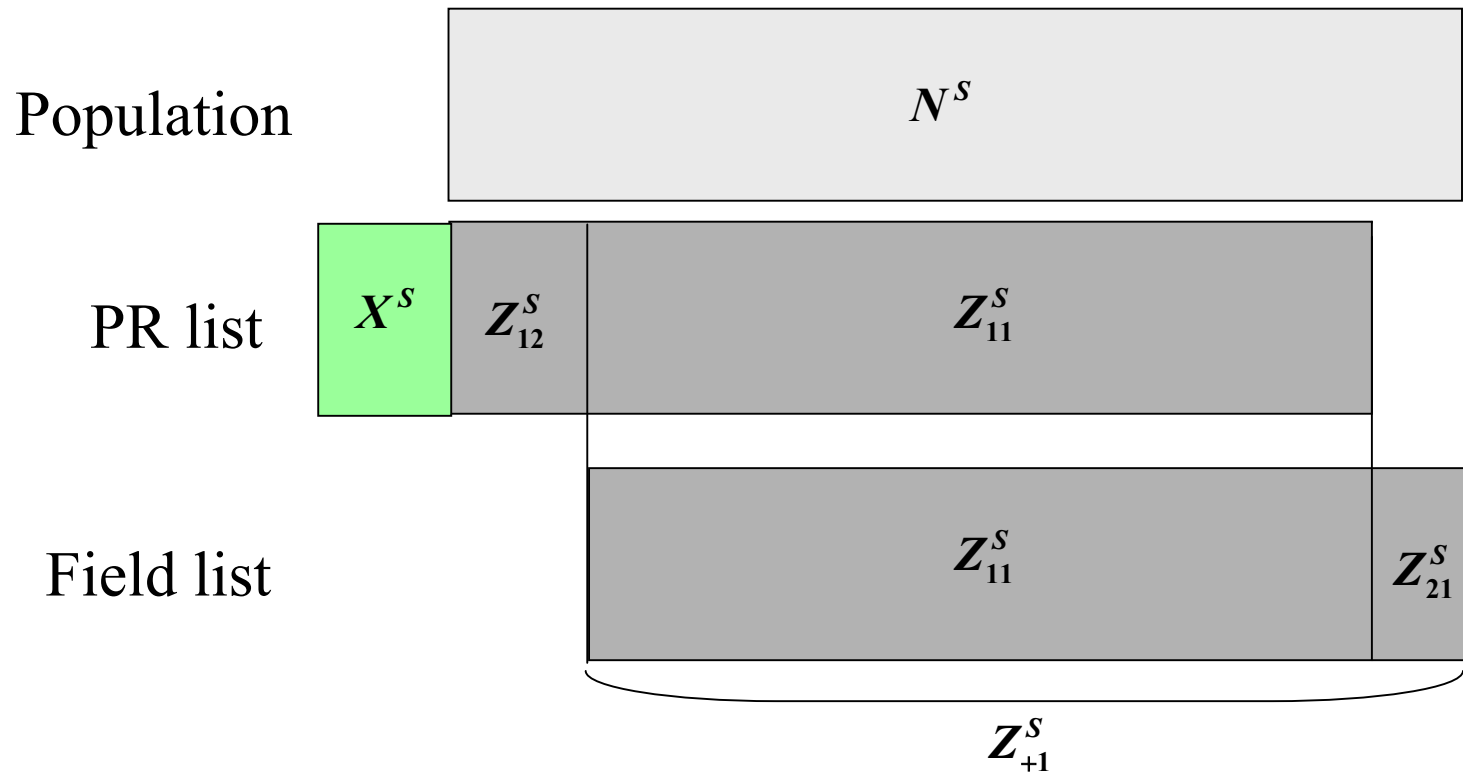
---

- Enumerators search the sampled EA`s in an attempt to enumerate all the households and all the people within households.
- The list generated by the field enumeration is matched with the PR list using an automated record linkage process.

Matching is done nationally, using ID numbers and other variables such as names and age.

# Coverage samples – U sample:

---



## Coverage survey - U sample

---

	In	Out	Total
In	$Z_{11}^S$	$Z_{12}^S$	$Z_{1+}^S$
Out	$Z_{21}^S$	$Z_{22}^S$	$Z_{2+}^S$
Total	$Z_{+1}^S$	$Z_{+2}^S$	$N^S$

- The undercoverage parameter is estimated by

$$\hat{p}_{1+} = Z_{11} / Z_{+1}$$

- Independence between field enumeration and the PR data is kept to avoid bias.

## Cover Survey - O sample

---

- In order to estimate the overcoverage parameter we must locate the people who were not enumerated at their PR address in order to determine where they actually live.
- ✓ At the first stage, the enumerators return to their EAs with list of names. Data is collected from the people themselves, relatives and neighbors.
- ✓ At the second stage, names are transferred to a CATI system, and information is collected by phones.

## O sample

---

- Using the information collected so far, we can determine  $X^S$  and  $Z_{1+}^S$ , and estimate the overcoverage parameter by

$$\hat{\lambda} = \frac{X^S}{\hat{N}^S} = \frac{X^S}{Z_{1+}^S / \hat{p}_{1+}}$$

( Recall that  $\lambda = EX / N$  )

## Population estimation:

---

- The population of each statistical area is divided into **estimation groups** (strata), which are homogeneous as much as possible with respect to the likelihood of being subject to coverage errors.
- Separate direct estimates of the two coverage parameters are calculated for each of the estimation groups within a statistical area.

## Population estimation:

---

- Define a census weight :

$$\hat{w} = \frac{1}{\hat{p}_{1+} + \hat{\lambda}}$$

- A census weight is assigned to every record in the PR, according to its estimation group.  
The weight reflects the number of people the record represents in the population.
- The census estimate for “any” population group is the sum of the PR weights assigned to its members.

## Beit-Shemesh pilot:

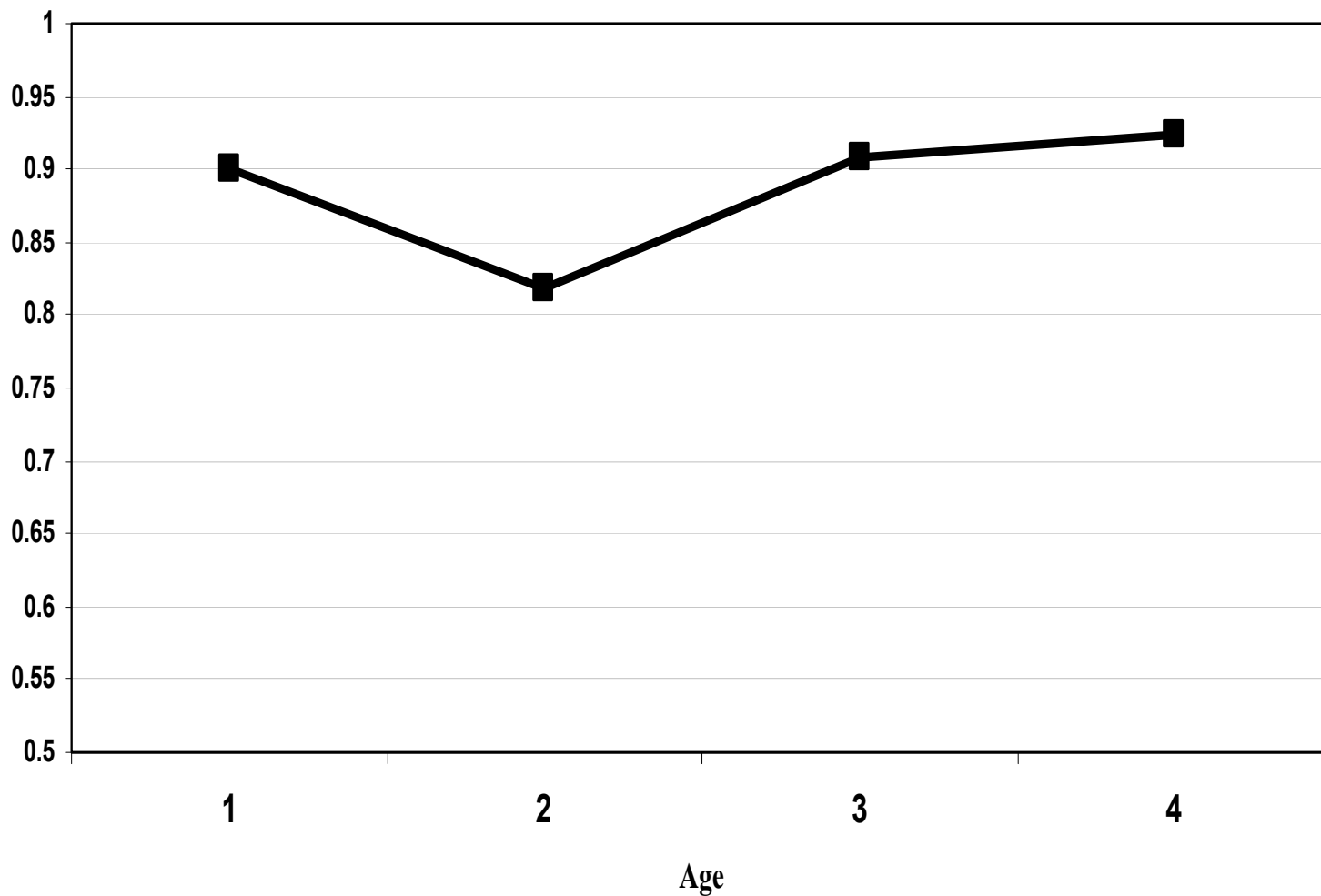
---

- The Integrated Census paradigm was first implemented in May 2002 in Beit-Shemesh, a town near Jerusalem.  
The town population is around 50,000 inhabitants divided into 13 statistical areas.
- Analyzing the results of the 2002 test, we found that among the PR variables, age has the strongest correlation with coverage errors.  
No differences were found between men and women.

*Undercount Rates by Estimation Group*

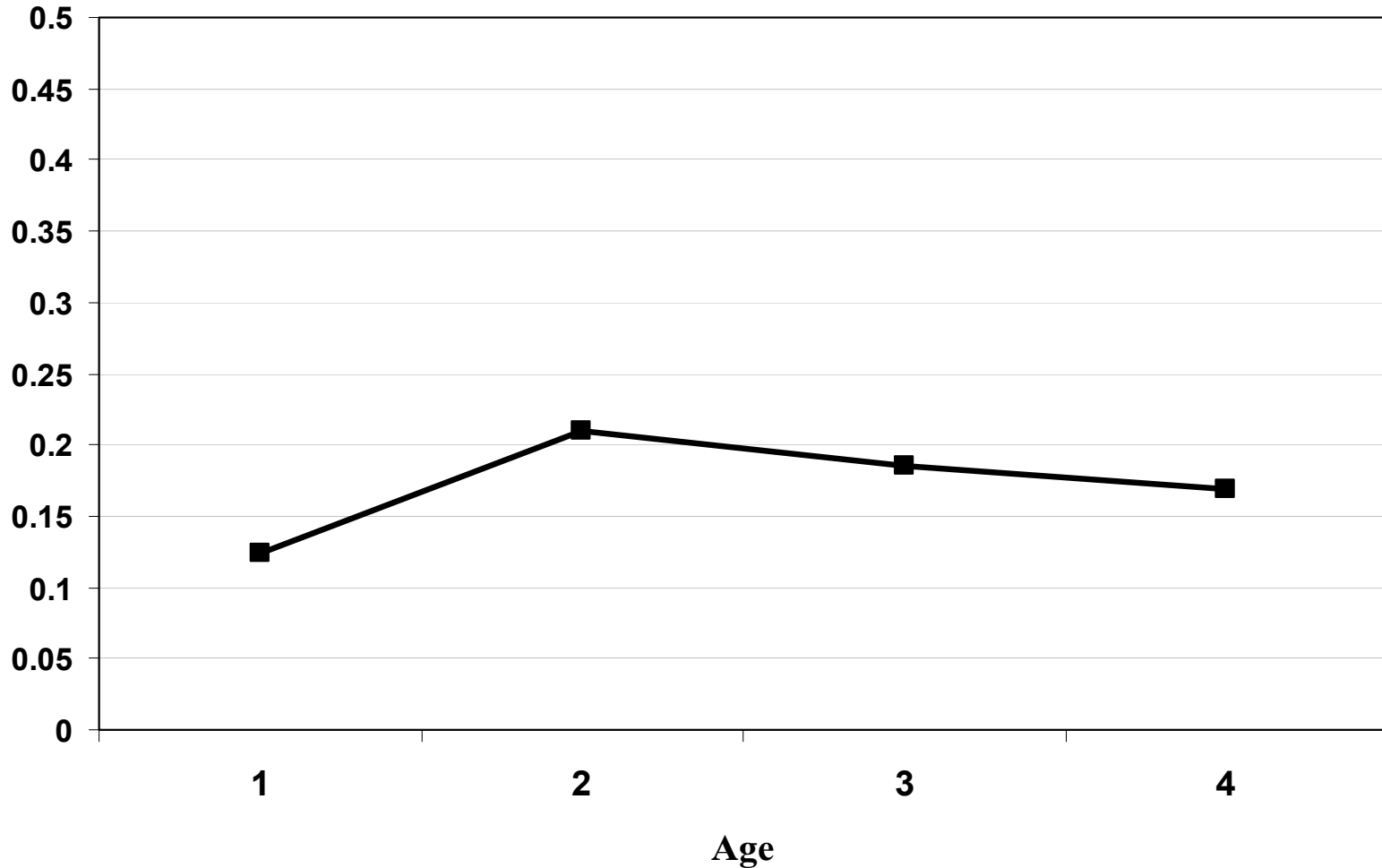
*Average over 13 statistical areas*

$\hat{p}_{1+}$

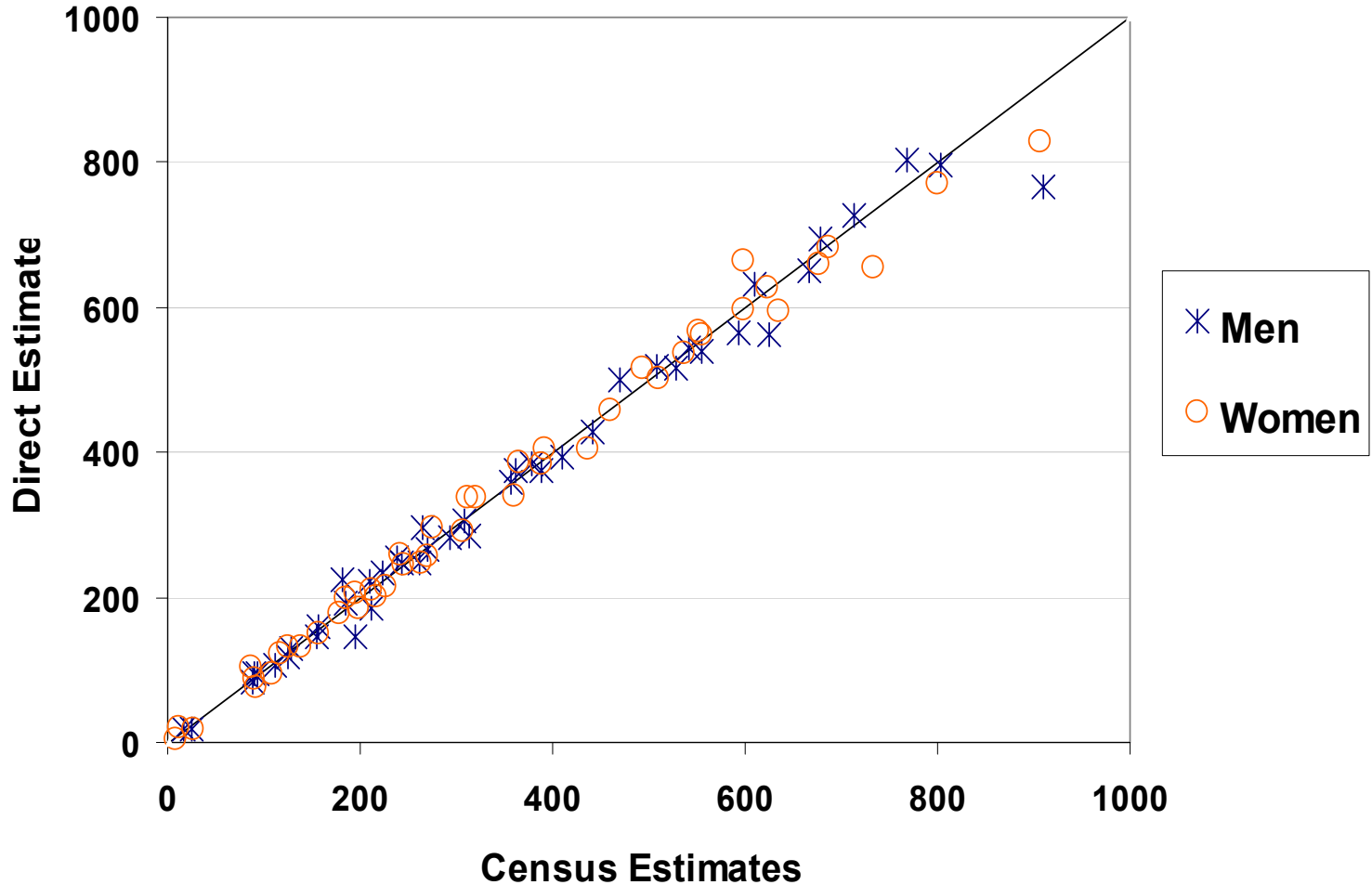


*Overcount Rates by Estimation Group*  
*Average over 13 statistical areas*

$\hat{\lambda}$



# Census Estimates vs Direct Estimates by Gender and Estimation Groups



# The Integrate Census process

Create an administrative file; geo-code all addresses



Design coverage surveys and collect data in the field



Link administrative and field data



Estimate coverage parameters and compute census weights



Evaluate estimates



*Thank you!*