# Stat472/572 Sampling: Theory and Practice


Instructor: Yan Lu

University of New Mexico

## Chapter 6: Sampling with Unequal Probabilities

# Motivation

Example 6.1: O'Brien et al. (1995)
A survey of nursing home residents in Philadelphia aimed to determine preferences regarding life-sustaining treatments.

- The study involved 294 nursing homes with a total of 37,652 beds (the number of residents was not known at the planning stage).
- Cluster sampling was used. Suppose an SRS of the 294 nursing homes is selected, followed by an SRS of 10 residents from each chosen home.
  —— A nursing home with 20 beds has the same probability of being selected as a nursing home with 1,000 beds.
  —— However, 10 residents from a 20-bed home represent fewer people than 10 residents from a 1,000-bed home.

**The above procedure results in a non-self-weighted sample. Possible design alternatives:**

- A one-stage cluster sample.
- A two-stage cluster design: an SRS of nursing homes followed by an equal proportion SRS of residents in each selected home.
- If an SRS is used at the first stage, $t_i$ is expected to be proportional to the number of beds in nursing home $i$, leading to estimators with large variance.

## The Study

- A sample of 57 nursing homes is drawn with probabilities proportional to the number of beds.
- An SRS of 30 beds (and their occupants) is then taken from a list of all beds within each selected nursing home.
- Each bed has an equal probability of being included in the sample (note the distinction between beds and occupants):

$$\begin{aligned} p &= \frac{\text{number of beds at the nursing home}}{\text{total beds in all nursing homes}} \\ &\times \frac{30}{\text{number of beds in the nursing home}} \\ &= \frac{30}{\text{total beds in all nursing homes}} \end{aligned}$$

- The same number of interviews is conducted at each nursing home, ensuring the cost is known before selecting the sample.
- This design likely results in estimators with smaller variance.

# Unequal Probabilities

- $\pi_i$ represents the probability that unit $i$ is selected as part of the sample.
- Most designs studied so far assume equal probabilities for $\pi_i$ across all units.
- In general, $\pi_i$ can vary with $i$, allowing for more flexible sampling designs.
- Sampling with unequal probabilities can yield significant advantages:
  - Decreases variances without the need for explicit stratification.
  - Allows deliberate variation in the selection probabilities of different primary sampling units (psus).
  - Compensates for unequal probabilities by applying appropriate weights during estimation.

# Sampling One PSU

- As a special case, we consider selecting just one ($n = 1$) of the $N$ primary sampling units (psus) to be included in the sample.
- Let the total for psu $i$ be denoted by $t_i$.
- Our goal is to estimate the population total, $t$.
- $\psi_i = p$(select unit $i$ on first draw).
- $\pi_i = p$(unit $i$ in sample).
- In the case of sampling one psu ($n = 1$), $\pi_i = \psi_i$.

**Example:** A town has four supermarkets, ranging in size from 100 square meters ($m^2$) to 1000 $m^2$. We want to estimate the total amount of sales in the four stores for last month by sampling just one of the stores.

- This is an illustration. We took a census, the total sales of the four supermarkets are 300 thousands.
- Expect a larger store would have more sales than a smaller store.
- The variability in total sales among several 1000 $m^2$ stores will be greater than the variability in total sales among several $100m^2$ stores.
- The probability that a store is selected on the first draw ($\psi_i$) is the same as the probability that the store is included in the sample ($\pi_i$).

| Store | Size ($m^2$) | $\psi_i$ | $t_i$(in Thousands) | $\hat{t}_\psi = \frac{t_i}{\psi_i}$ | $(\hat{t}_\psi - t)^2$ |
|---|---|---|---|---|---|
| A | 100 | 1/16 | 11 | 176 | 15,376 |
| B | 200 | 2/16 | 20 | 160 | 19,600 |
| C | 300 | 3/16 | 24 | 128 | 29,584 |
| D | 1000 | 10/16 | 245 | 392 | 8,464 |
| Total | 1600 | 1 | 300 | | |

$$
\begin{aligned}
E[\hat{t}_\psi] &= \sum p(S)\hat{t}_\psi s \\
&= \frac{1}{16}(176) + \frac{2}{16}(160) + \frac{3}{16}(128) + \frac{10}{16}(392) \\
&= 300
\end{aligned}
$$

$$
\begin{aligned}
V[\hat{t}_\psi] &= E[(\hat{t}_\psi - t)^2] \\
&= \sum p(S)(\hat{t}_{\psi S} - t)^2 \\
&= \sum_{i=1}^{N} \psi_i \left( \frac{t_i}{\psi_i} - t \right)^2 \\
&= \frac{1}{16}(15,376) + \frac{2}{16}(19,600) + \frac{3}{16}(29,584) \\
&+ \frac{10}{16}(8,464) \\
&= 14,248
\end{aligned}
$$

**Consider an SRS of size 1**

| Store | Size ($m^2$) | $\psi_i$ | $t_i$(in Thousands) | $\hat{t}_\psi = t_i/\psi_i = 4t_i$ | $(\hat{t}_\psi - t)^2$ |
|---|---|---|---|---|---|
| A | 100 | 1/4 | 11 | 44 | 65,536 |
| B | 200 | 1/4 | 20 | 80 | 48,400 |
| C | 300 | 1/4 | 24 | 96 | 41,616 |
| D | 1000 | 1/4 | 245 | 980 | 462,400 |
| Total | 1600 | 1 | 300 | | |

- Probability of selecting each unit is $\psi_i = 1/4$
- SRS estimator is unbiased
- $V(\hat{t}_{srs}) = 154,488 \gg V[\hat{t}_\psi] = 14,248$
  ——The variance of SRS is significantly larger compared to that of unequal probability sampling.

# One-Stage Sampling with Replacement ($n > 1$)

- $\psi_i = p$(select unit (psu) $i$ on first draw)
  $= p$(select unit $i$ on any given draw)
  —- The probability of selecting unit $i$ on the first draw is the same as its probability on any subsequent draw.
- $\pi_i = p$(unit $i$ appears in the sample)
  —- This implies $\pi_i = 1 - (1 - \psi_i)^n$, where $n$ is the number of draws.
- $Q_i =$ number of times unit (psu) $i$ is included in the sample
  —- The total number of draws satisfies $\sum_{i=1}^{N} Q_i = n$, and the expected number of times unit $i$ is selected is $E(Q_i) = n\psi_i$.

- **Estimator of the Total:**

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}$$

  - In sampling with replacement, we obtain $n$ independent estimates of the population total, one for each unit included in the sample.
  - The final estimate is the average of these $n$ independent estimates.

- Unbiasedness Proof:

$$\begin{aligned} E(\hat{t}_\psi) &= \frac{1}{n}\sum_{i=1}^{N} E(Q_i)\frac{t_i}{\psi_i} \\ &= \frac{1}{n}\sum_{i=1}^{N} n\psi_i \frac{t_i}{\psi_i} \\ &= \sum_{i=1}^{N} t_i \\ &= t \end{aligned}$$

**Variance:**

$$V(\hat{t}_\psi) = \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2$$

$$\begin{aligned}\hat{V}(\hat{t}_\psi) &= \frac{1}{n}\cdot\frac{1}{n-1}\sum_{i\in\Re}\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{N}Q_i\frac{\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2}{n-1}\end{aligned}$$

where $\Re$ denote the set of $n$ units in the sample, including the repeats.

$$E[\hat{V}(\hat{t}_\psi)] = V(\hat{t}_\psi)(\text{ see proof on page 227})$$

**Estimator of the mean:**

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{\hat{M}_{0\psi}},$$

where $\hat{M}_{0\psi} = \frac{1}{n}\sum_{i\in\Re}\frac{M_i}{\psi_i}$.

$$\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{(\hat{M}_{0\psi})^2}\cdot\frac{1}{n}\cdot\frac{1}{n-1}\sum_{i\in\Re}\left(\frac{t_i}{\psi_i} - \frac{\hat{\bar{y}}_\psi M_i}{\psi_i}\right)^2$$

# Probability proportional to size (pps) sampling

- Many totals in a psu are related to the number of elements in the psu
  —let $M_i$ be the size of psu $i$
  —let $M_0 = \sum_{i=1}^{N} M_i$ be the size of the population
- Take $\psi_i = M_i/M_0$
  —a large psu has a greater chance of being in the sample than a small psu.

# One-Stage pps sampling

- $\frac{t_i}{\psi_i} = t_i \frac{M_0}{M_i} = M_0 \bar{y}_i$
- Estimator of the total: $\hat{t}_\psi = \frac{1}{n} \sum_{i \in \Re} M_0 \bar{y}_i$
- Estimator of the population mean: $\hat{\bar{y}}_\psi = \frac{1}{n} \sum_{i \in \Re} \bar{y}_i$,

  ——this is the average of the sampled psu means
- $\hat{M}_{0\psi} = M_0$ for every possible sample
- Variance of the estimated mean:
  $\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i \in \Re} (\bar{y}_i - \hat{\bar{y}}_\psi)^2$
- The pps estimators can be computed by treating the sampled psu means ($\bar{y}_i$) as individual observations, then calculating their mean and sample variance directly.

## Example: Estimating the Total Number of Physicians in the United States

The file `statepop.dat` contains data from an unequal-probability sample of 100 counties in the United States. Our goal is to estimate the total number of physicians in the country.

- Counties were selected using the cumulative-size method from the listings in the *City and County Data Book* (1994), with probabilities proportional to their populations:

$$\psi_i = \frac{M_i}{M_0}$$

  where $M_i$ is the population size of county $i$ and $M_0$ is the total population of all counties.

- Sampling was conducted with replacement.
- Very large counties appear multiple times in the sample. For instance, Los Angeles County, the most populous in the United States, appears four times.
- Since larger counties tend to have more physicians, probability-proportional-to-size (pps) sampling is expected to be effective for estimating the total number of physicians.

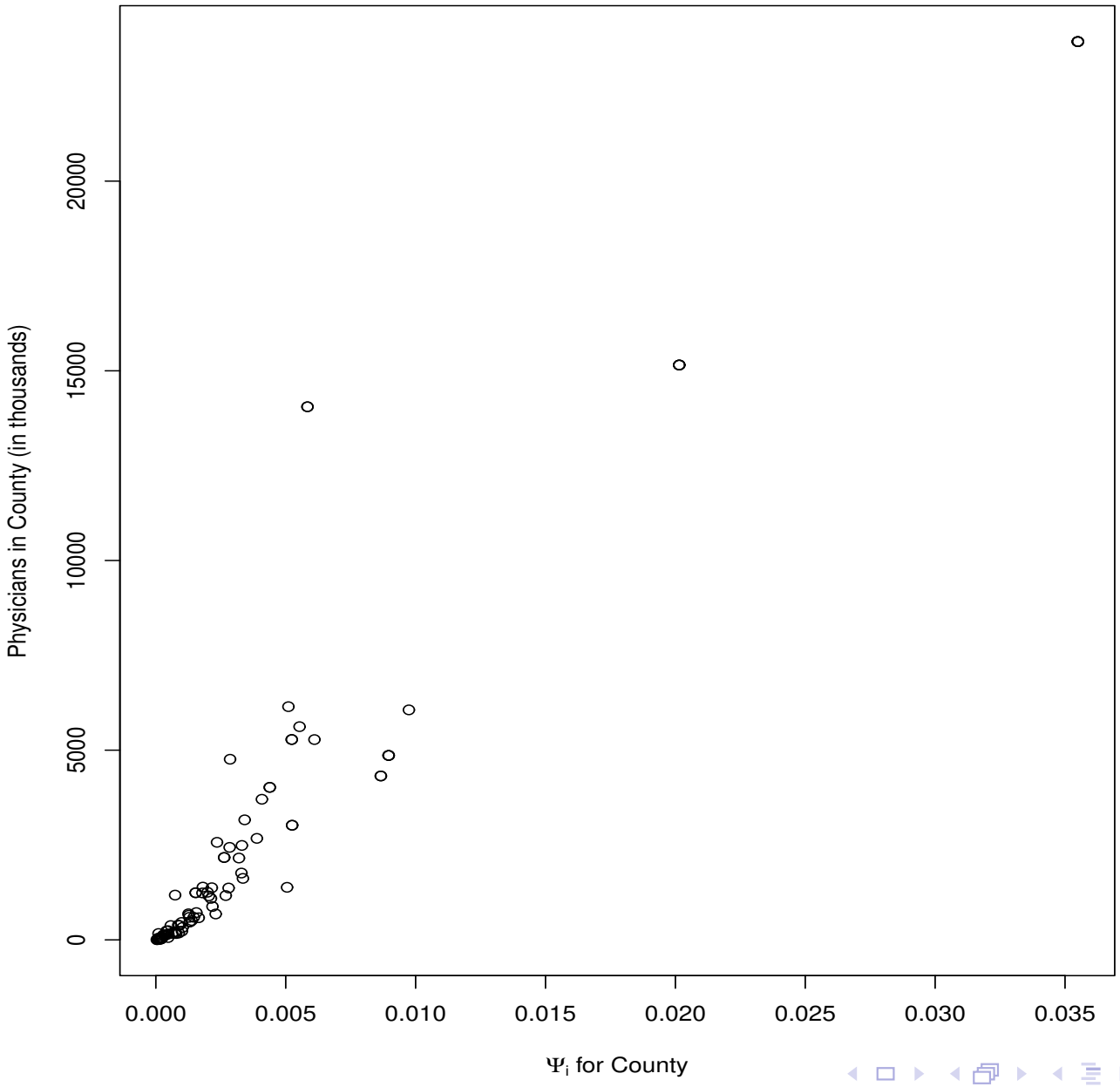
Physicians in County (in thousands)
0
5000
10000
15000
20000
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Ψi for County

**TABLE 6.3**
Sampled Counties in Example 6.5

| State | County | Population Size, $M_i$ | $\psi_i$ | Number of Physicians, $t_i$ | $\frac{t_i}{\psi_i}$ |
|---|---|---|---|---|---|
| AL | Wilcox | 13,672 | 0.00005360 | 4 | 74,627.72 |
| AZ | Maricopa | 2,209,567 | 0.00866233 | 4320 | 498,710.81 |
| AZ | Maricopa | 2,209,567 | 0.00866233 | 4320 | 498,710.81 |
| AZ | Pinal | 120,786 | 0.00047353 | 61 | 128,820.64 |
| AR | Garland | 76,100 | 0.00029834 | 131 | 439,095.36 |
| AR | Mississippi | 55,060 | 0.00021586 | 48 | 222,370.54 |
| CA | Contra Costa | 840,585 | 0.00329541 | 1761 | 534,379.68 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| VA | Chesterfield | 225,225 | 0.00088297 | 181 | 204,990.72 |
| WA | King | 1,557,537 | 0.00610613 | 5280 | 864,704.59 |
| WI | Lincoln | 27,822 | 0.00010907 | 28 | 256,709.47 |
| WI | Waukesha | 320,306 | 0.00125572 | 687 | 547,096.42 |
| | | | | | |
| | | average | | | 570,304.30 |
| | | std. dev. | | | 414,012.30 |

an idea of the spread involved in the population estimates, and may help you identify unusual psus (Figure 6.1b).

The sample was chosen using the cumulative-size method; Table 6.3 shows the sampled counties arranged alphabetically by state. The $\psi_i$'s were calculated using $\psi_i = M_i/M_0$. The average of the $t_i/\psi_i$ column is 570,304.3, the estimated total number of physicians in the United States. The standard error of the estimate is $414{,}012.3/\sqrt{100} = 41{,}401$. For comparison, the *County and City Data Book* lists a total of 532,638 physicians in the United States; a 95% CI using our estimate includes the true value.

These estimates can be found using the SAS code on the website. Partial output is given below:

```
    Data Summary

Number of Observations          100
Sum of Weights            2450.71956

                    Statistics

                                Std Error
```

- The unusual observation is New York County, New York

| | Landarea square miles | Population Size | Number of Physicians |
|---|---|---|---|
| New York County | 28 | 1,489,066 | 14,052 |
| Mean for the 100 selected counties | 1188.04 | 1,048,753 | 2,979.87 |

- The estimated total number of physicians in the United States is $570,304$
- The standard error of the estimate is 41401.23
- For comparison, the *City and Country Data Book* lists a total of $532,638$ physicians in the United States, a value that is less than 1 SE away from the estimate

# Two-Stage Sampling with Replacement

- The key difference between two-stage sampling with replacement and one-stage sampling with replacement is that in two-stage sampling, we must estimate $t_i$ (the total for primary sampling unit, PSU $i$).
- If PSU $i$ appears in the sample more than once, there are $Q_i$ estimates of the total for PSU $i$: $\hat{t}_{i1}, \hat{t}_{i2}, \ldots, \hat{t}_{iQ_i}$.
- The estimator for the total is:

$$\hat{t}_\psi = \frac{1}{n} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \frac{\hat{t}_{ij}}{\psi_i}$$

- The variance of the estimator is:

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \frac{\left( \frac{\hat{t}_{ij}}{\psi_i} - \hat{t}_\psi \right)^2}{n-1}$$

## Estimator of the Population Mean

- The estimator of the population mean is:

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{\hat{M}_{0\psi}},$$

where

$$\hat{M}_{0\psi} = \frac{1}{n}\sum_{i\in\mathcal{W}} \frac{M_i}{\psi_i}.$$

- Variance of the estimated mean:

$$\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{(\hat{M}_{0\psi})^2} \cdot \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^{N}\sum_{j=1}^{Q_i} \left(\frac{\hat{t}_{ij}}{\psi_i} - \frac{\hat{\bar{y}}_\psi M_i}{\psi_i}\right)^2.$$

**pps sampling is a special case** when the probability of selecting each primary sampling unit (psu) is proportional to its size, i.e.,

$$\psi_i = \frac{M_i}{M_0},$$

where $M_i$ is the size of psu $i$, and $M_0$ is the total size of all psus in the population.

## Example 6.7: Average Number of Legs of Puppies

- Goal: Estimate the average number of legs of healthy puppies in Sample City's puppy homes.
- Sample City has two puppy homes:
  - Puppy Palace: 30 puppies
  - Dog's Life: 10 puppies
- Sampling Design: pps sampling
  - Puppy homes are selected with probabilities proportional to the size of each puppy home:

$$\text{P(Puppy Palace)} = \frac{30}{30+10} = \frac{3}{4}, \quad \text{P(Dog's Life)} = \frac{10}{30+10} = \frac{1}{4}.$$

—-this ensures larger homes (with more puppies) have a higher chance of being selected
—-pps sampling is effective because it accounts for the variation in home sizes, potentially reducing the variance of the estimator
  - After a home is selected, a simple random sample (SRS) of 2 puppies is taken from the chosen home.

| | $M_i$ | $\psi_i$ | $m_i$ | $\hat{t}_i$ | $\hat{t}_\psi = \frac{\hat{t}_i}{\psi_i}$ | $\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{M_0}$ |
|---|---|---|---|---|---|---|
| Puppy Palace | 30 | 3/4 | 2 | 30*4=120 | 160 | 4 |
| Dog's Life | 10 | 1/4 | 2 | 10*4=40 | 160 | 4 |

- Either possible sample results in an estimated average of $\hat{\bar{y}}_\psi = 160/40 = 4$ legs per puppy
- The variance of the estimator is zero

## Notes: Sampling with Replacement

- Sampling with replacement is advantageous because:
  - It is simple and straightforward to select the sample.
  - Estimators for both the population total. mean and their variance are easy to compute.
- However, if the population size ($N$) is small:
  - Sampling with replacement becomes less efficient compared to many sampling designs without replacement.
  - This is because repeated selections of the same unit reduce the effective sample size, increasing the variance of the estimators.

# Weights in Unequal probability sampling with replacement

**One stage sampling with replacement with only one psu**

- $\pi_i = \psi_i$
- $w_{ij} = w_i = 1/\psi_i$
- $\hat{t}_\psi = \sum_{i \in \mathcal{W}} \sum_{j=1}^{M_i} w_{ij} y_{ij}$
- $\hat{\bar{y}}_\psi = \dfrac{\sum_{i \in \mathcal{W}} \sum_{j=1}^{M_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{W}} \sum_{j=1}^{M_i} w_{ij}}$

**One Stage Sampling with replacement: general Case (more than one psu)**

$\pi_i \approx n\psi_i$, the larger the population, the closer of $\pi_i$ to $n\psi_i$

$$
\begin{aligned}
w_i &= \frac{1}{\text{expected number of hits}} \\
&= \frac{1}{E[Q_i]} \\
&= \frac{1}{n\psi_i}
\end{aligned}
$$

- $w_{ij} = w_i = \dfrac{1}{n\psi_i}$
- $\hat{t}_\psi = \sum_{i \in \mathcal{W}} \sum_{j=1}^{M_i} w_{ij} y_{ij}$
- $\hat{\bar{y}}_\psi = \dfrac{\sum_{i \in \mathcal{W}} \sum_{j=1}^{M_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{W}} \sum_{j=1}^{M_i} w_{ij}}$
- $\psi_i$s are unequal, the sample is not self-weighting
- In one-stage pps sampling, elements in large psus have smaller weights than elements in small psus

**Two stage cluster sampling with replacement**

Sampling weight for element sampled from psu $i$

$$w_i = \frac{1}{n\psi_i}\frac{M_i}{m_i}$$

pps sampling

- $\psi_i = \dfrac{M_i}{M_0}$
- $w_i = \dfrac{M_0}{nM_i}\dfrac{M_i}{m_i} = \dfrac{M_0}{nm_i}$
- If $m_i$s are the same within each psu, sample is self-weighting
  —-equal interviewer works loads
  —-sample size of ssu's known in advance
  —-if $t_i \propto \psi_i$, more efficient

## Review: SRS without replacement

- $\pi_i = p(\text{unit } i \text{ in the sample}) = n/N$
- $\sum_{i=1}^{N} \pi_i = n$
- $\pi_{ij} = p(\text{unit i and j in the sample}) = \frac{n(n-1)}{N(N-1)}$
- $\sum_{i=1, j\neq i}^{N} \pi_{ij} = n(n-1)$

  Proof: Let

$$Z_i = \begin{cases} 1 & \text{if unit} \quad i \in S \\ 0 & \text{otherwise} \end{cases}$$

$p(z_i = 1) = \pi_i$

$E(z_i) = 1 * p(z_i = 1) = \pi_i$

$$\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} E(z_i) = E\left(\sum_{i=1}^{N} z_i\right) = n$$

$$
\begin{aligned}
\sum_{i=1,j\neq i}^{N} \pi_{ij} &= \sum_{i=1,j\neq i}^{N} p(Z_i = 1, Z_j = 1) \\
&= \sum_{i=1,j\neq i}^{N} E(Z_i Z_j) \\
&= \sum_{i=1}^{N} E[Z_i(n - Z_i)] \\
&= \sum_{i=1}^{N} (E(nZ_i) - E(Z_i^2)) \\
&= \sum_{i=1}^{N} \left( n \cdot \frac{n}{N} - \frac{n}{N} \right) \\
&= n(n-1)
\end{aligned}
$$

**Example 6.8: Supermarket Example**

| Store | Size ($m^2$) | $\psi_i$ | $t_i$(in Thousands) |
|---|---|---|---|
| A | 100 | 1/16 | 11 |
| B | 200 | 2/16 | 20 |
| C | 300 | 3/16 | 24 |
| D | 1000 | 10/16 | 245 |
| Total | 1600 | 1 | 300 |

- Want to select two psus without replacement and with unequal probabilities.
  —Recall $\psi_i = p$(select unit $i$ on first draw)

$$p(\text{store } A \text{ chosen on first draw}) = \psi_A = 1/16$$

$$\begin{aligned} & p(\text{store } B \text{ chosen on second draw} | A \text{ chosen on first draw}) \\ = & \frac{\frac{2}{16}}{1 - \frac{1}{16}} \\ = & \frac{\psi_B}{1 - \psi_A} \end{aligned}$$

In general,

$$\begin{aligned} & p(\text{unit } i \text{ chosen first, unit } k \text{ chosen second}) \\ = & p(\text{unit } i \text{ chosen first}) p(\text{ unit } k \text{ chosen second} | \text{ unit } i \text{ chosen first}) \\ = & \psi_i \frac{\psi_k}{1 - \psi_i} \end{aligned}$$

$$
\begin{aligned}
& p(\text{unit } k \text{ chosen first, unit } i \text{ chosen second}) \\
= \quad & \psi_k \frac{\psi_i}{1-\psi_k}
\end{aligned}
$$

$$
\begin{aligned}
& p(\text{units } i \text{ and } k \text{ in sample}) \\
= \quad & \pi_{ik} = \psi_i \frac{\psi_k}{1-\psi_i} + \psi_k \frac{\psi_i}{1-\psi_k}
\end{aligned}
$$

$$
\begin{aligned}
\pi_{AB} &= \psi_A \frac{\psi_B}{1-\psi_A} + \psi_B \frac{\psi_A}{1-\psi_B} \\
&= \frac{1}{16} \cdot \frac{\frac{2}{16}}{1-\frac{1}{16}} + \frac{2}{16} \cdot \frac{\frac{1}{16}}{1-\frac{2}{16}} \\
&= 0.0173
\end{aligned}
$$

The probability that psu $i$ is in the sample $\pi_i = \sum_{S:i \in S} P(S)$.

```
supermarket<-data.frame(store=c('A','B','C','D'),
area=c(100,200,300,1000), ti=c(11,20,24,245))
supermarket$psi<-supermarket$area/sum(supermarket$area)
psii<-supermarket$area/sum(supermarket$area)
piik<- psii %*% t(psii/(1-psii)) +
(psii/(1-psii)) %*% t(psii)
diag(piik)<-rep(0,4) # diagonal entries: zero
piik  # joint inclusion probabilities
pii<-apply(piik,2,sum)
pii # inclusion probabilities
```

```
> cbind(piik,pii)
        A           B           C           D         pii
[1,] A 0.0000000 0.01726190 0.02692308 0.1458333 0.1900183
[2,] B 0.0172619 0.00000000 0.05563187 0.2976190 0.3705128
[3,] C 0.0269230 0.05563187 0.00000000 0.4567308 0.5392857
[4,] D 0.1458333 0.29761905 0.45673077 0.0000000 0.9001832
```

**TABLE 6.5**
Inclusion probabilities ($\pi_i$) and joint inclusion probabilities ($\pi_{ik}$) for samples of size 2 that could be selected using the method in Example 6.8. The entries of the table are the $\pi_{ik}$'s for each pair of stores (rounded to four decimal places); the margins give the $\pi_i$'s for the four stores

| | | Store $k$ | | | | |
|---|---|---|---|---|---|---|
| | | A | B | C | D | $\pi_i$ |
| Store $i$ | A | — | 0.0173 | 0.0269 | 0.1458 | 0.1900 |
| | B | 0.0173 | — | 0.0556 | 0.2976 | 0.3705 |
| | C | 0.0269 | 0.0556 | — | 0.4567 | 0.5393 |
| | D | 0.1458 | 0.2976 | 0.4567 | — | 0.9002 |
| | $\pi_k$ | 0.1900 | 0.3705 | 0.5393 | 0.9002 | 2.0000 |

size 2 consists of psus $i$ and $k$:

$$\text{For } n = 2, \quad P(\text{units } i \text{ and } k \text{ in sample}) = \pi_{ik} = \psi_i \frac{\psi_k}{1 - \psi_i} + \psi_k \frac{\psi_i}{1 - \psi_k}.$$

The probability that psu $i$ is in the sample is then

$$\pi_i = \sum_{\mathcal{S}: i \in \mathcal{S}} P(\mathcal{S}).$$

Table 6.5 gives the $\pi_i$'s and $\pi_{ik}$'s for the supermarkets. ■

## 6.4.1 The Horvitz–Thompson Estimator for One-Stage Sampling

Assume we have a without-replacement sample of $n$ psus, and we know the **inclusion probability**

$$\pi_i = P(\text{unit } i \text{ in sample})$$

# Unequal probability sampling without replacement (one-stage)

- $\pi_i = p(\text{unit } i \text{ in sample})$
- $\pi_i/n$ is the average probability that a unit will be selected on one of the draws: the probability we would assign to the $i$th unit's being selected on draw $k(k = 1, \cdots, n)$ if we did not know the true probabilities
- the estimator $t_i/\psi_i$ is then estimated by $t_i/(\pi_i/n)$

# Horvitz-Thompson (HT) Estimator for One-Stage Sampling

Horvitz and Thompson 1952

$$\begin{aligned}\hat{t}_{\mathrm{HT}} &= \frac{1}{n}\sum_{i\in S}\frac{t_i}{\pi_i/n} \\ &= \sum_{i\in S}\frac{t_i}{\pi_i} \\ &= \sum_{i=1}^{N} Z_i\frac{t_i}{\pi_i}\end{aligned}$$

Unbiased: $E[\hat{t}_{\mathrm{HT}}] = \sum_{i=1}^{N} \pi_i \frac{t_i}{\pi_i} = t$

Variance:

$$\begin{aligned} V[\hat{t}_{\mathrm{HT}}] &= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N} \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1, k \neq i}^{N} (\pi_i \pi_k - \pi_{ik}) \left( \frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2 \end{aligned}$$

Horvitz-Thompson (HT) Estimator

$$\hat{V}_1[\hat{t}_{\mathrm{HT}}] = \sum_{i\in S}(1-\pi_i)\frac{t_i^2}{\pi_i^2} + \sum_{i\in S}\sum_{k\in S, k\neq i}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\cdot\frac{t_i}{\pi_i}\cdot\frac{t_k}{\pi_k}$$

Sen-Yates-Grundy Estimator

$$\hat{V}_2[\hat{t}_{\mathrm{HT}}] = \frac{1}{2}\sum_{i\in S}\sum_{k\in S, k\neq i}\frac{\pi_i\pi_k-\pi_{ik}}{\pi_{ik}}\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2$$

**Example 6.9**, The HT estimator for a sample of 2 supermarkets in Example 6.8 with joint inclusion probabilities given in Table 6.7.

- To select the first psu, we generate a random integer from $1, \cdots, 16$, the random integer we generate is 12, which tells us that store $D$ is selected on the first draw.
- We then remove the values $7, \cdots, 16$ corresponding to store $D$, and generate a second random integer from $1, \cdots, 6$, we generate 6, which tells us to select store $C$ on the second draw.

```
> supermarket2<-supermarket[3:4,]
# these are the unit inclusion probs when n=2
> supermarket2$pii <- pii[3:4]
# joint probability matrix for stores C and D
> jointprob<-piik[3:4,3:4]
# set diagonal entries equal to pii
> diag(jointprob)<-supermarket2$pii
> jointprob
          [,1]      [,2]
[1,] 0.5392857 0.4567308
[2,] 0.4567308 0.9001832
```

```
> dht<- svydesign(id=~1,  fpc=~pii, data=supermarket2,
+                       pps=ppsmat(jointprob),variance="HT")
> dht
Sparse-matrix design object:
 svydesign(id = ~1, fpc = ~pii, data = supermarket2,
pps = ppsmat(jointprob),
    variance = "HT")
> svytotal(~ti,dht)
    total     SE
ti 316.67 82.358
```

```
> dsyg<- svydesign(id=~1,  fpc=~pii, data=supermarket2,
+                        pps=ppsmat(jointprob),variance="YG")
> dsyg
Sparse-matrix design object:
 svydesign(id = ~1, fpc = ~pii, data = supermarket2,
pps = ppsmat(jointprob),
    variance = "YG")
> svytotal(~ti,dsyg)
    total     SE
ti 316.67 57.094
```

- Given the sample $\{C, D\}$, the HT estimate of the total sales is:

$$\hat{t}_{HT} = \sum_{i \in S} \frac{t_i}{\pi_i} = \frac{24}{0.5393} + \frac{245}{0.9002} = 316.6639$$

- Variance and standard error of the HT estimate:

$$\hat{V}(\hat{t}_{HT}) = 6782.8, \quad SE(\hat{t}_{HT}) = 82.358$$

- SYG (Sen-Yates-Grundy) approximation for variance and standard error:

$$\hat{V}_{SYG}(\hat{t}_{HT}) = 3259.8, \quad SE_{SYG}(\hat{t}_{HT}) = 57.094$$

**Example: Basu's Elephant**

Debabrata Basu (1971) famously critiqued the Horvitz-Thompson (HT) estimator in his essay *"An Essay on the Logical Foundations of Survey Sampling, Part I"*. He illustrated his point with the following story:

- A circus owner plans to ship 50 elephants and needs an estimate of their total weight.
- She decides to weigh only one elephant, Sambo (a middle-sized elephant), and uses $50 \cdot y_{\text{Sambo}}$ (where $y_{\text{Sambo}}$ is Sambo's weight) as the total weight estimate.
- The circus statistician is horrified because this method gives zero probability of sampling the other elephants.
- To address this, the statistician proposes a different plan:
  - Assign a 99% probability of selecting Sambo.
  - Assign the remaining 1% probability equally among the other 49 elephants.

## The Sampling Results

As expected, Sambo is selected.

- The circus owner asks, "Since Sambo was selected, isn't the total weight estimate just $50 \cdot y_{\text{Sambo}}$?"
- The statistician replies, No, the HT estimate is:

$$\hat{t}_{HT} = \frac{y_{\text{Sambo}}}{0.99} = 1.01 \cdot y_{\text{Sambo}}$$

- The owner then asks, "What if the largest elephant, Jumbo, had been selected instead?"
- The statistician explains, The HT estimate would then be:

$$\hat{t}_{HT} = \frac{y_{\text{Jumbo}}}{\frac{0.01}{49}} = 4900 \cdot y_{\text{Jumbo}}$$

- Upon hearing this, the circus owner immediately fires the statistician!

**What's Wrong with Basu's Elephant Example?**

The example highlights several critical issues with this approach:

- **Sample size is too small**: Only one elephant is selected, making the estimate unreliable.
- **Selection probabilities are too extreme**: The probabilities heavily favor Sambo, making the design inherently biased and prone to large variability.
- **Huge standard error**: Extreme selection probabilities lead to large variability in the HT estimator.
- **Unstable estimates**: While the HT estimator is unbiased in theory (over repeated samples, the average estimate is close to the truth), individual estimates can deviate drastically, especially with such extreme designs.

# The Horvitz-Thompson (HT) Estimator for Two-stage Sampling

$$\hat{t}_{\text{HT}} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}$$

If $E(\hat{t}_i) = t_i$, $E[\hat{t}_{\text{HT}}] = t$
Variance Estimator (HT form)

$$\hat{V}_1[\hat{t}_{\text{HT}}] = \sum_{i \in S} (1-\pi_i) \frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i \in S} \sum_{k \in S, k \neq i} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k} + \sum_{i \in S} \frac{\hat{V}(\hat{t}_i)}{\pi_i}$$

Variance Estimator (SYG form)

$$\hat{V}_2[\hat{t}_{\text{HT}}] = \frac{1}{2} \sum_{i \in S} \sum_{k \in S, k \neq i} \frac{\pi_i \pi_k - \pi_{ik}}{\pi_{ik}} \left( \frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_k}{\pi_k} \right)^2 + \sum_{i \in S} \frac{\hat{V}(\hat{t}_i)}{\pi_i}$$

# Recommendation: Variance Estimation Using With-Replacement Sampling

For most situations, we recommend using the with-replacement sampling variance estimator to avoid potential instability and computational complexity.

$$\begin{aligned}\hat{V}_{\text{WR}}(\hat{t}_{\text{HT}}) &= \frac{1}{n}\cdot\frac{1}{n-1}\sum_{i\in S}\left(\frac{n\hat{t}_i}{\pi_i}-\hat{t}_{\text{HT}}\right)^2 \\ &= \frac{n}{n-1}\sum_{i\in S}\left(\frac{\hat{t}_i}{\pi_i}-\frac{\hat{t}_{\text{HT}}}{n}\right)^2\end{aligned}$$

**Notes:**

- In practice, samples are often drawn without replacement, but the variance is calculated assuming with-replacement sampling.
- Using the with-replacement estimator generally results in a larger variance than the true variance (overestimation), which is conservative and ensures robustness.

## Unequal Probability Sampling in Practice

- Many government surveys rely on unequal probability sampling techniques.
- Stratification is often used first to reduce the variation in probabilities ($\pi_i$).

## Example: Random Digit Dialing (RDD)

- Construct a frame using area codes and prefixes (e.g., 505-243-).
- Draw a random sample of suffixes (0000-9999), dial numbers, and check if they are residential. If not, discard.

For psu $i$, let $M_i$ be the number of residential numbers:

$$\begin{aligned} p(\text{dial number}) &= p(\text{psu } i \text{ selected}) \times p(\text{number selected in psu } i) \\ &= \frac{M_i}{\sum_{j=1}^{N} M_j} \cdot \frac{m_i}{M_i} \end{aligned}$$

If all $m_i$'s are equal, the sampling becomes self-weighting.

# Disclaimer

Slides are intended for a course based on the book: *Sampling: Design and Analysis, Third Edition* by Sharon L. Lohr
All examples, datasets, and pages directly scanned and included in this chapter are from the textbook.
References to papers or books cited in these slides can be found in the Bibliography section of the textbook.