# Aspects spéciaux de l'échantillonnage et de l'estimation

Pierre Duchesne

August 1, 2017

# Estimation par domaines (SSW, Chap. 10; Satin et Shastry, p. 53

- On a déjà discuté de différentes sous-populations reliées à l'échantillonnage.
- Parfois ces sous-populations sont nommées des domaines.
- Exemples de domaines: âge, sexe, activité, profession.

# Estimation dans chaque domaine

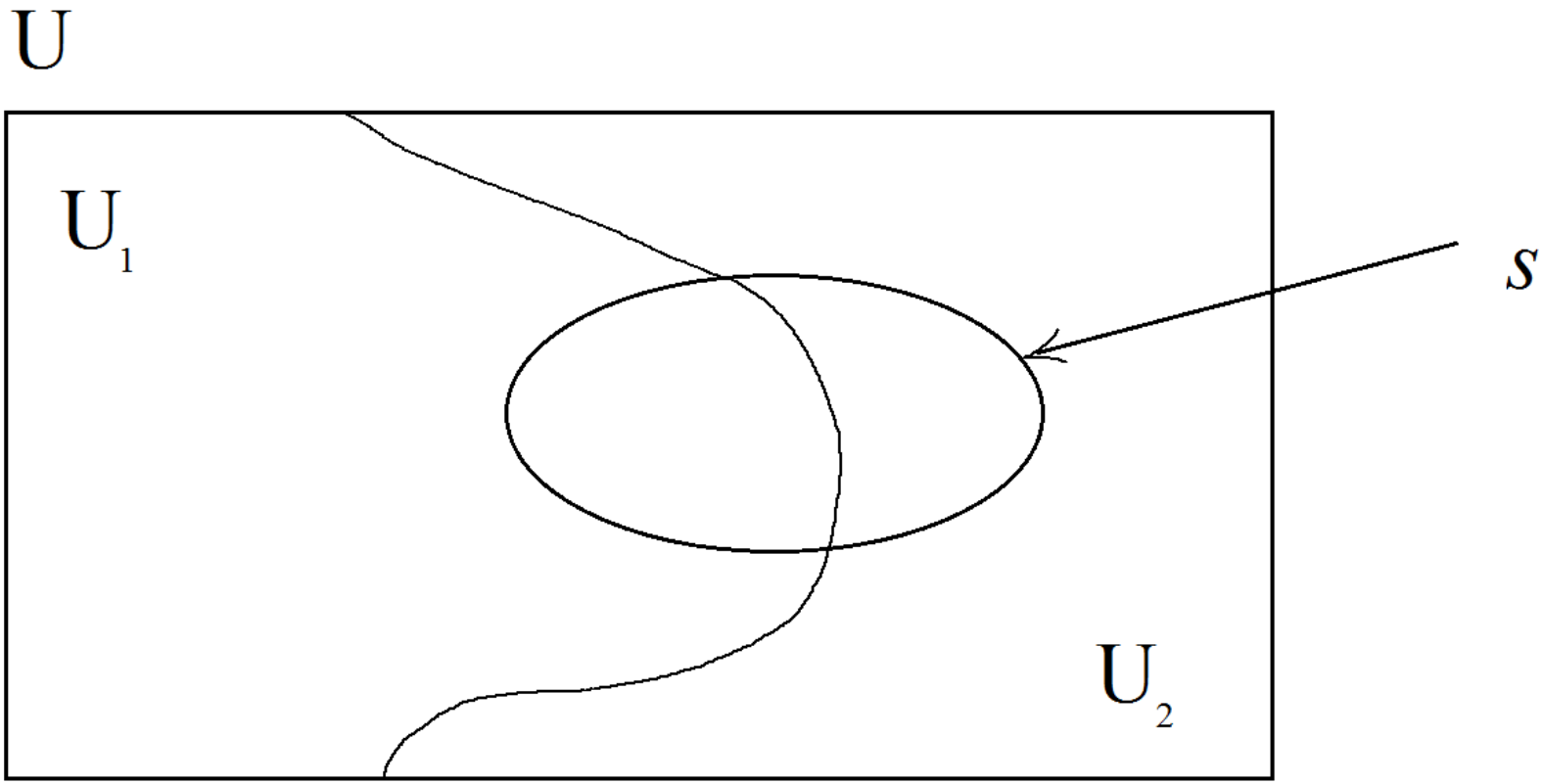
- En pratique, l'estimation dans chaque domaine nécessite un échantillon assez grand.
- Idéalement:
  1. On s'arrange pour stratifier selon ces domaines. Dans un tel cas, les strates sont exactement les domaines.
  2. On peut planifier, comme dans un plan stratifié, la taille des échantillons dans chaque strate (domaine).
- Quand on ne peut procéder de façon idéale, on doit utiliser l'échantillon $s$ à notre disposition.
  1. L'utilisation de la théorie des domaines est utile lors de questions qui surviennent après le sondage.

# Exemple

Considérons:

$U_1$ : population des hommes,

$U_2$ : population des femmes.



# Estimation dans chaque domaine (suite)

- Dans l'exemple précédent, l'échantillon $s$ pourrait être sélectionné selon les plans SI, SY ou BE dans la population $U$.
- Exemple: Ainsi, sous un plan SI, si l'échantillon est de taille $n = 1000$, dans une population comprenant autant d'hommes que de femmes, on s'attend d'obtenir 500 femmmes et 500 hommes dans $s$.
- Cependant, il peut en être différemment en pratique compte tenu de l'aspect aléatoire de la sélection de $s$.

# Estimation dans chaque domaine (suite)

- On va élaborer la théorie entourant l'estimation dans chaque domaine.
- On peut considérer l'estimation de totaux et de moyennes. L'estimation de la moyenne n'est pas aussi simple que cela pourrait paraître à prime abord.
- On va considérer le cas de deux domaines mais l'approche se généralise facilement.

# Problème des grandes unités (Satin et Shastry, p. 53)

- Dans les enquêtes sur les populations humaines, il arrive fréquemment que quelques grandes villes aient une influence marquuée.
  - Beaucoup de petites villes, quelques très grandes villes.
- Il en est de même dans les enquêtes agricoles, les enquêtes portant sur les entreprises.
  - Quelques unités (fermes, entreprises) ont une influence considérable.
  - Revenues de certaines entreprises présentent des valeurs très élevées par rapport aux valeurs généralement observées (PME versus multinationales).

# Que se passe-t-il si le plan d'échantillonnage est mal conçu?

Dans une enquête portant sur les revenues des entreprises, les estimations des caractéristiques d'intérêts risquent d'être:

- Très élevées si les grosses unités sont sur-représentées;
- Très faibles si les grosses unités sont sous-représentées.

# Solution idéale

Idéalement, que faire pour surmonter les difficultés dans de telles populations?
Dans la mesure du possible, on peut tenter de stratifier les entreprises (ou les fermes, ou les grosses unités) en fonction de leur taille.

- On pourrait créer une strate "prendre tout". On pourrait choisir toutes les grandes unités en donnant $\pi_k = 1$ si $k$ est une grande entreprise, disons pour les unités $k$ dans $U_{grande}$.
- Pour les petites entreprises, disons la population $U_{petite}$, on pourrait échantillonner (disons par un plan tel le plan SI) normalement.

# Si ce n'est pas possible?

- L'estimateur Horvitz-Thompson est une méthode d'estimation.
- D'autres techniques sont possibles.
- Le problème des grandes unités est relié au problème des valeurs aberrantes.
- La théorie des méthodes robustes peut s'avérer utile.
- Une technique d'estimation possible est la méthode du *M*-estimateur.

# *M*-estimateur

- Pour illustrer la technique, limitons-nous au plan SI, avec $\pi_k = n/N$.
- On peut remarquer que l'estimateur Horvitz-Thompson est obtenu comme solution de l'équation:

$$\sum_s (y_k - \mu) = 0$$

- L'équation précédente est une *équation d'estimation*.
- Le *M*-estimateur est obtenu comme solution de l'équation:

$$\sum_s \psi(y_k - \mu) = 0$$

# *M*-estimateur (suite)

- Il faut choisir la fonction $\psi(\cdot)$.
- Un exemple est la fonction de Huber qui est définie comme suit:

$$\psi(x) = \begin{cases} -c & x < -c, \\ x & x \in [-c, c], \\ c & x > c. \end{cases}$$

- Bien que cela dépasse un premier cours d'échantillonnage, il est possible d'étudier les propriétés de biais et de variance de ces estimateurs, de considérer des plans d'échantillonnages plus généraux que SI, etc.