Shapley Decomposition

European Doctoral School of Demography

19 May 2022

Benjamin Elbers

benjamin.elbers@nuffield.ox.ac.uk

Nuffield College
UNIVERSITY OF OXFORD

LEVERHULME
CENTRE
FOR
DEMOGRAPHIC
SCIENCE

Overview

1. Introduction to decomposition
2. Shapley decomposition
3. Applications in demography
4. Advanced Topics: Many factors and hierarchical factors

Organization

- GitHub repository
- Exercise sheet available there as RMarkdown file
- Additional exercises to try later

Part 1: Introduction to Decomposition

Why decomposition?

- Loosely defined: understand a statistic (often a difference) through its constituting factors
- In *demography*: understanding the differences (e.g., in terms of death rates, fertility behavior) between two populations—either over time or across areas
- In *machine learning*: understanding the prediction of a complex model
- In *statistics*: splitting the variance into a within and a between part: $\mathrm{Var}[X] = \mathrm{E}[\mathrm{Var}[X \mid Y]] + \mathrm{Var}[\mathrm{E}[X \mid Y]]$
- In *sociology*: understand the factors that contribute to an increase/decrease in segregation

A classic

- Kitagawa (1955): “When comparing the incidence of some phenomenon in two or more groups, social researchers place much emphasis on the need for holding constant those related factors that would tend to distort the comparison. For example, before comparing the death rates for the residents of two areas, demographers frequently control the factors of differences between the areas in age, sex and race composition.”

Example data (Clogg and Eliason 1998)

- “Would you like to have another child?”, by no. of children

	4+ children		One child	
	N	%	N	%
20 to 24	27	37.0	363	90.1
25 to 29	152	19.1	208	76.9
30 to 34	224	15.2	96	56.2
35 to 39	239	5.0	59	20.3
40 to 44	211	6.2	48	10.4
Overall	**853**	**11.5**	**774**	**72.1**

- The overall rate is a function of both the age distribution and the age-specific rates.

Notation

- Let's introduce some notation:
  - y_{ij} is the rate for the ith group in the jth population,
  - n_{ij} is the size of the ith group in the jth population, and
  - $N_j = \sum_j n_{ij}$ is the size of population j.

Notation

	4+ children		One child	
	N	%	N	%
20 to 24	n_{11}	y_{11}	n_{12}	y_{12}
25 to 29	n_{21}	y_{21}	n_{22}	y_{22}
30 to 34	n_{31}	y_{31}	n_{32}	y_{32}
35 to 39	n_{41}	y_{41}	n_{42}	y_{42}
40 to 44	n_{51}	y_{51}	n_{52}	y_{52}
Overall	N_1	$\bar{y}_1 = \sum_{i=1}^{5} p_{i1} y_{i1}$	N_2	$\bar{y}_2 = \sum_{i=1}^{5} p_{i1} y_{i1}$

- Let $p_{ij} = n_{ij}/N_j$ such that $\sum_i p_{ij} = 1$.
- $\bar{y}_j$ is the overall rate in population j.

Standardization

- One approach is standardization: Remove the compositional effect by using the identical distribution for both overall rates.
- "What would the overall rate of the women with one child be if this population had the same distribution as the women with 4+ children?"
  - Answer: $\bar{y}_{\text{standardized}(1)} = \sum_{i=1}^{5} p_{i1} y_{i2}$
- "What would the overall rate of the women with 4+ children be if this population had the same distribution as the women with one child?"
  - Answer: $\bar{y}_{\text{standardized}(2)} = \sum_{i=1}^{5} p_{i2} y_{i1}$

Standardization

	4+ children		One child		Difference
	N	%	N	%	
20 to 24	27	37.0	363	90.1	
25 to 29	152	19.1	208	76.9	
30 to 34	224	15.2	96	56.2	
35 to 39	239	5.0	59	20.3	
40 to 44	211	6.2	48	10.4	
Overall	**853**	**11.5**	**774**	**72.1**	**60.6**
Standardized(1)		**11.5**		**39.6**	**28.1**
Standardized(2)		**25.1**		**72.1**	**47.0**

Standardization

- Standardization requires a “reference” distribution—this can also be an artificial distribution.
- For instance, we could use

  $$\hat{p}_i = \frac{p_{i1} + p_{i2}}{2}$$

  as the identical reference distribution for both populations.
- Different reference distributions will give different answers.

Decomposition

- While standardization is useful, Kitagawa (1955) introduced an even more useful idea: **decomposition**.
- The idea is to decompose the difference into two parts:

$$\bar{y}_2 - \bar{y}_1 = \underbrace{\text{(differences in composition)}}_{\Delta p} + \underbrace{\text{(differences in rates)}}_{\Delta y}$$

Kitagawa (1955)

- She proposed

$$\Delta p = \sum_i \tfrac{y_{i1}+y_{i2}}{2}\left(p_{i2} - p_{i1}\right)$$
$$\Delta y = \sum_i \tfrac{p_{i1}+p_{i2}}{2}\left(y_{i2} - y_{i1}\right)$$

such that

$$\bar{y}_2 - \bar{y}_1 = \Delta p + \Delta y$$

Exercises – Part 1

1. Proof $\bar{y}_2 - \bar{y}_1 = \Delta p + \Delta y$.
2. How does the term Δy relate to standardization?
3. In R, use the `children` dataset to decompose the difference in overall rates using Kitagawa's method.