Heterogeneity, variance, and factorial variance components

Hal Caswell
Silke van Daalen

Institute for Biodiversity and Ecosystem Dynamics
University of Amsterdam

Wittgenstein Center Conference 2023
Exploring Population Heterogeneities

European Research Council
Established by the European Commission

Heterogeneity and variance

- what's coming
  - variance partitioning
  - a new method: multiple factors and their interactions
  - some examples: longevity, sex, race, location
  - new ways of thinking about the mixing distribution, and why it is your friend

Heterogeneous populations are mixtures

Population

$$\overbrace{\begin{array}{rcl} \text{group} & = & 1 \\ \text{proportion} & = & \pi_1 \\ E(\xi|\text{group=1}) & = & m_1 \\ V(\xi|\text{group = 1}) & = & v_1 \end{array} \cdots \Longleftrightarrow \cdots \begin{array}{rcl} \text{group} & = & G \\ \text{proportion} & = & \pi_G \\ E(\xi|\text{group=G}) & = & m_G \\ V(\xi|\text{group = G}) & = & v_G \end{array}}^{\textbf{Population}} \quad (0)$$

π = **mixing distribution**

Probability distribution of ξ in the population is a *mixture* of the distributions in each group

Heterogeneous populations are mixtures

Population

$$
\begin{aligned}
E(\xi) &= E_\pi(m) \\
V(\xi) &= E_\pi(v) + V_\pi(m)
\end{aligned}
\qquad (0)
$$

$$
\begin{aligned}
\text{group} &= 1 \\
\text{proportion} &= \pi_1 \\
E(\xi|\text{group=1}) &= m_1 \\
V(\xi|\text{group = 1}) &= v_1
\end{aligned}
\quad \cdots \Longleftrightarrow \cdots \quad
\begin{aligned}
\text{group} &= G \\
\text{proportion} &= \pi_G \\
E(\xi|\text{group=G}) &= m_G \\
V(\xi|\text{group = G}) &= v_G
\end{aligned}
\qquad (0)
$$

Variance partitioning: within and between groups[1]

$$V(\xi) \quad = \quad \underbrace{E_\pi \left[V(\xi|\text{group}) \right]}_{\text{within-group}} + \underbrace{V_\pi \left[E(\xi|\text{group}) \right]}_{\text{between-group}}$$

- within-group = stochasticity
- between-group = heterogeneity
- variance ratio = contribution of heterogeneity

$$\mathcal{K} = \frac{V_{\text{between}}}{V_{\text{within}} + V_{\text{between}}}$$

[1] see Caswelll 2023, The contributions of stochastic demography and social inequality to lifespan variability. Demographic Research 49: 309-354.

Demographic outcomes

ξ = longevity. Moments from Markov chains (Feichtinger 1971, Caswell 2001, 2009)

ξ = lifetime fertility. Moments from Markov chain with rewards (Caswell 2011, van Daalen and Caswell 2015, 2017)

ξ = "healthy" longevity. Moments (prevalence and incidence) from Markov chains with rewards (Caswell and Zarulli 2018, Zarulli and Caswell 2022, Caswell and van Daalen 2021)

But . . . how many factors?

- most studies look at one factor at a time[2]
- combinations of multiple factors: variance is due to:
  - each factor *and*
  - all interactions between factors
- method available for multi-factor studies
  - any number of factors
  - any number of levels
  - any demographic outcome ξ that has means and variances

[2] We might see an example with 4 simultaneous factors, see Bergeron-Boucher et al., **Session 5 on Thursday**.

Mixture distributions extended to multiple (two) factors

Factor A at N_A levels, factor B at N_B levels.

Variance components:

$$
\begin{aligned}
V(\xi) &= V_{\text{within}} + \underbrace{V_{\text{between}}}_{V_A + V_B + V_{AB}} \\
V_A &= \text{variance due to A} \\
V_B &= \text{variance due to B} \\
V_{AB} &= \text{variance due to interaction}
\end{aligned}
$$

The key is to partition V_{between} using the multi-factor mixing distribution

Mixing distribution must be one of two types[3]

flat. All probabilities equal. Corresponds to a well-designed experiment. All factor combinations are evaluated equally in their contributions to variance.

rank-one. Also called proportional. The mixing distribution can be assembled from its marginals, and the mixture weights are proportional across factors A and B.

These distributions ask different questions.

[3] A well known issue in experimental design.

$$\mathbf{m}_B = \sum_i \sum_k (\mathbf{M} \circ \mathbf{\Pi}) \oslash \sum_i \sum_k \mathbf{\Pi} \qquad N_B \times 1$$

$$\mathbf{m}_C = \sum_i \sum_j (\mathbf{M} \circ \mathbf{\Pi}) \oslash \sum_i \sum_j \mathbf{\Pi} \qquad N_C \times 1$$

$$\mathbf{m}_{AB} = \sum_k (\mathbf{M} \circ \mathbf{\Pi}) \oslash \sum_k \mathbf{\Pi} \qquad N_A \times N_B$$

$$\mathbf{m}_{AC} = \sum_j (\mathbf{M} \circ \mathbf{\Pi}) \oslash \sum_j \mathbf{\Pi} \qquad N_A \times N_C$$

$$\mathbf{m}_{BC} = \sum_i (\mathbf{M} \circ \mathbf{\Pi}) \oslash \sum_i \mathbf{\Pi} \qquad N_B \times N_C$$

$$\mathbf{m}_{ABC} = (\mathbf{M} \circ \mathbf{\Pi}) \oslash \mathbf{\Pi} \qquad N_A \times N_B \times N_C$$

$$\boldsymbol{\pi}_A = \sum_j \sum_k \mathbf{\Pi} \qquad N_A \times 1$$

$$\boldsymbol{\pi}_B = \sum_i \sum_k \mathbf{\Pi} \qquad N_B \times 1$$

$$\boldsymbol{\pi}_C = \sum_i \sum_j \mathbf{\Pi} \qquad N_C \times 1$$

$$\boldsymbol{\pi}_{AB} = \sum_k \mathbf{\Pi} \qquad N_A \times N_B$$

$$\boldsymbol{\pi}_{AC} = \sum_j \mathbf{\Pi} \qquad N_A \times N_C$$

$$\boldsymbol{\pi}_{BC} = \sum_i \mathbf{\Pi} \qquad N_B \times N_C$$

$$\boldsymbol{\pi}_{ABC} = \mathbf{\Pi} \qquad N_A \times N_B \times N_C$$

$$V_{\mathrm{A}} = \mathbb{V}(\mathbf{m}_A, \boldsymbol{\pi}_A)$$

$$V_{\mathrm{B}} = \mathbb{V}(\mathbf{m}_B, \boldsymbol{\pi}_B)$$

$$V_{\mathrm{C}} = \mathbb{V}(\mathbf{m}_A, \boldsymbol{\pi}_A)$$

$$V_{\mathrm{AB}} = \mathbb{V}(\mathrm{vec}\,\mathbf{m}_{AB}, \mathrm{vec}\,\boldsymbol{\pi}_{AB}) - V$$

$$V_{\mathrm{AC}} = \mathbb{V}(\mathrm{vec}\,\mathbf{m}_{AC}, \mathrm{vec}\,\boldsymbol{\pi}_{AC}) - V$$

$$V_{\mathrm{BC}} = \mathbb{V}(\mathrm{vec}\,\mathbf{m}_{BC}, \mathrm{vec}\,\boldsymbol{\pi}_{BC}) - V$$

$$V_{\mathrm{ABC}} = \mathbb{V}(\mathrm{vec}\,\mathbf{m}_{ABC}, \mathrm{vec}\,\boldsymbol{\pi}_{ABC}) -$$

Alas, no time for algebra

A two-factor example

U.S. Longevity 2020: sex×race[4]

- 5 racial/ethnic groups, 2 sexes
- means for each factor combination

	Means				
	Hispanic	NHAIAN	NHA	NHB	NHW
male	74.6	63.8	81.1	67.8	74.8
female	81.3	70.7	85.9	75.3	80.1

- variances for each combination

	Variances				
	Hispanic	NHAIAN	NHA	NHB	NHW
male	289	408	218	372	294
female	219	391	163	315	235

[4] United States Life Tables 2020, NVSS 71(1).

U.S. Longevity: sex×race

Mixing distributions

- flat

$$\pi_{\text{race}} = \begin{pmatrix} \text{Hispanic} \\ \text{NHAIAN} \\ \text{NHA} \\ \text{NHB} \\ \text{NHW} \end{pmatrix} = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix} \qquad \pi_{\text{sex}} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}.$$

- rank-one, based on estimated population by race

$$\pi_{\text{race}} = \begin{pmatrix} \text{Hispanic} \\ \text{NHAIAN} \\ \text{NHA} \\ \text{NHB} \\ \text{NHW} \end{pmatrix} = \begin{pmatrix} 0.19 \\ 0.01 \\ 0.06 \\ 0.13 \\ 0.62 \end{pmatrix} \qquad \pi_{\text{sex}} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}.$$

Mixing distributions: sex×race

Mixing distributions **ask different questions**

- flat

$$\pi_{\text{race}} = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}$$

- what are the contributions of sex and race, as such?
- each sex-race combination is equally valued
- "experimental" interpretation

- rank-one, based on estimated population by race

$$\pi_{\text{race}} = \begin{pmatrix} 0.19 \\ 0.01 \\ 0.06 \\ 0.13 \\ 0.62 \end{pmatrix}$$

- contributions of sex and race
- in a population with the composition π_{race}
- "survey" interpretation[a]
- 81% Hispanic and NHW; don't care much about NHAIAN

[a] Kendall and Stuart 1976. The Advanced Theory of Statistics, Vol. 3

Variance components: U.S. sex×race

Flat mixing

Component	Variance
Race	31.5
Sex	9.7
Race × sex	0.3
(between-group)	41.5
Stochasticity	290.3
Total	331.8
$\mathcal{K}$	0.125

Rank-one mixing

Component	Variance
Race	7.4
Sex	8.5
Race × sex	0.2
(between-group)	16.1
Stochasticity	269.3
Total	285.4
$\mathcal{K}$	0.056

A three-factor example

Sex, race, and U.S. state of residence[5]

- differences in life expectancy among U.S. states (73–80y) are comparable to differences due to race and sex
- life table data
  - 40 U.S. states
  - two races ("white" and "black")
  - two sexes
- 3 main effects, 3 two-way interactions, 1 three-way interaction

[5] U.S. Decennial Life Tables 1999–2001, NVSS 60(9)

Variance components: sex×race×U.S. state

Flat mixing

Component	Variance
A=sex	9.23
B=race	7.50
C=state	1.37
AB=sex × race	0.061
AC=sex × state	0.209
BC=race × state	0.567
ABC=sex × race × state	0.170
(between-groups)	19.11
Stochasticity	303.04
Total	322.14
$\mathcal{K}$	0.059

Conclusions

1. Heterogeneous populations are mixtures
2. Variance from stochasticity and heterogeneity
3. Now possible to partition variance due to heterogeneity
   - contributions of multiple factors,
   - and interactions,
   - longevity, lifetime fertility, healthy longevity, more
4. Longevity example: sex, race, state
   - heterogeneity still makes small contribution
   - interactions not important
5. Lifetime fertility may behave differently (consequences of failure)

Do you have data on multiple factors? Happy to talk about it.

Thank you

Longevity: sex and U.S. state of residence[6]

Flat mixing

Component	Variance
State	3.69
Sex	7.74
State $\times$ sex	0.088
(between-groups)	11.53
Stochasticity	275.6
Total	287.1
$\mathcal{K}$	0.040

Rank-one mixing

Component	Variance
State	2.88
Sex	7.93
State $\times$ sex	0.053
(between-groups)	10.86
Stochasticity	270.03
Total	280.9
$\mathcal{K}$	0.039

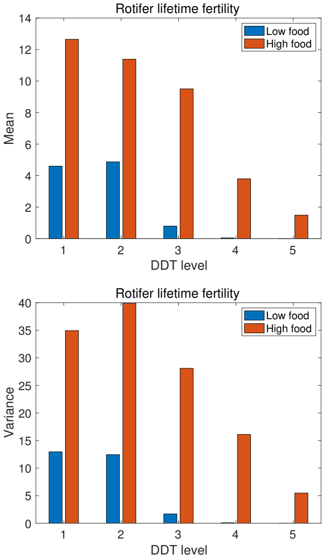
[6] U.S. State Life Tables, NVSS 71(2). Rank-one mixing based on total state population.

Lifetime fertility: nutrition and pollution

- Rotifers: cute little tiny aquatic organisms
- Model species for aging and toxicology studies
- Factorial design[a]
    - 2 levels of nutrition (high and low food)
    - 5 levels of environmental stress (DDT exposure)
- means and variances calculated from Markov chain with rewards

[a] Rao and Sarma (1986), Caswell (2001)

Lifetime fertility



Flat mixing

Component	Variance
A=Food	8.12
B=DDT	10.09
AB=Food $\times$ DDT	1.84
(between-group)	20.05
Stochasticity	15.2
Total	35.2
$\mathcal{K}$	0.57