

# Analysing phonological systems: on Bayesian typological research

George Moroz

Linguistic Convergence Laboratory, NRU HSE, Moscow, Russia

24 August 2019

Societas Linguistica Europaea, 52nd Annual Meeting, Leipzig University

Presentation is available here: [tinyurl.com/y3wtkcbq](https://tinyurl.com/y3wtkcbq)



## In this talk I will cover the following:

- Goals of linguistic typology
- Different strategies of sampling
- The Bayesian way of thinking about linguistic typology
- Case study: vowels

## Goals of linguistic typology

- Attest distributions (statistical and areal) of typological values

## Goals of linguistic typology

- Attest distributions (statistical and areal) of typological values
- Find a correlation between the distributions of different typological categories
  - Absolute universals
  - Distributional patterns and tendencies
  - Semantic maps
  - Diachronic change of typological values

## Goals of linguistic typology

- Attest distributions (statistical and areal) of typological values
- Find a correlation between the distributions of different typological categories
  - Absolute universals
  - Distributional patterns and tendencies
  - Semantic maps
  - Diachronic change of typological values
- Find a correlation between linguistic and non-linguistic patterns
  - Population movements
  - Population size
  - Language contact
  - Sociolinguistic parameters
  - Geopolitical environment (including the spread of diseases)

## Goals of linguistic typology

- Attest distributions (statistical and areal) of typological values
- Find a correlation between the distributions of different typological categories
  - Absolute universals
  - Distributional patterns and tendencies
  - Semantic maps
  - Diachronic change of typological values
- Find a correlation between linguistic and non-linguistic patterns
  - Population movements
  - Population size
  - Language contact
  - Sociolinguistic parameters
  - Geopolitical environment (including the spread of diseases)
- Deal with mixed typological values

## Frequentist typological research

- Formulate a theoretical problem

There is a category in some languages with values **VAL 1** and **VAL 2**.

## Frequentist typological research

- Formulate a theoretical problem  
There is a category in some languages with values **VAL 1** and **VAL 2**.  
What is the probability  $\theta$  of finding **VAL 1** in a randomly picked language?



## Frequentist typological research

- Formulate a theoretical problem  
There is a category in some languages with values **VAL 1** and **VAL 2**.  
What is the probability  $\theta$  of finding **VAL 1** in a randomly picked language?
- Get a grant, hire some students, or select a holiday you want to spend working on this topic...

## Frequentist typological research

- Formulate a theoretical problem  
There is a category in some languages with values **VAL 1** and **VAL 2**.  
What is the probability  $\theta$  of finding **VAL 1** in a randomly picked language?
- Get a grant, hire some students, or select a holiday you want to spend working on this topic...
- Pick a sample of languages, calculate the desired statistics, e. g.  $\hat{\theta}$

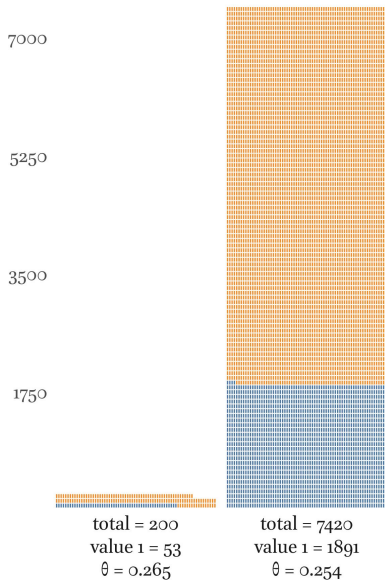
## Frequentist typological research

- Formulate a theoretical problem  
There is a category in some languages with values **VAL 1** and **VAL 2**.  
What is the probability  $\theta$  of finding **VAL 1** in a randomly picked language?
- Get a grant, hire some students, or select a holiday you want to spend working on this topic...
- Pick a sample of languages, calculate the desired statistics, e. g.  $\hat{\theta}$
- From now on  $\hat{\theta}$  is the best estimation of  $\theta$  that you know

## Frequentist typological research

- Formulate a theoretical problem  
There is a category in some languages with values **VAL 1** and **VAL 2**.  
What is the probability  $\theta$  of finding **VAL 1** in a randomly picked language?
- Get a grant, hire some students, or select a holiday you want to spend working on this topic...
- Pick a sample of languages, calculate the desired statistics, e. g.  $\hat{\theta}$
- From now on  $\hat{\theta}$  is the best estimation of  $\theta$  that you know, add some **confidence intervals** of you need to convince an editor who is mad about statistics
- After you have published your paper, your project is finished

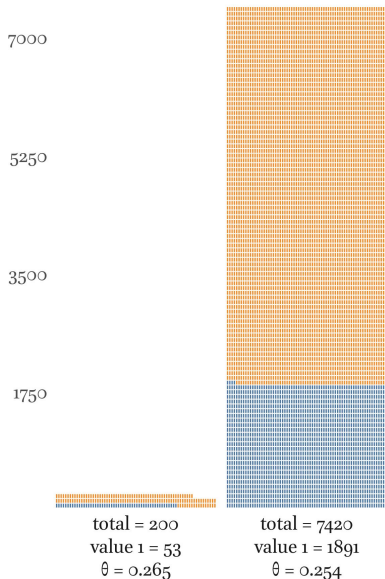
# There are different types of sampling



## Random sampling

each member of the population has an equal probability of selection

# There are different types of sampling



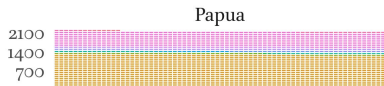
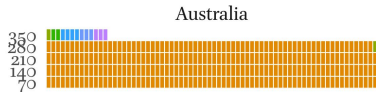
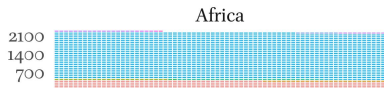
## Random sampling

each member of the population has an equal probability of selection

!!! but each language is grouped in a **language family** and an **area**, so observations are not independent...



## There are different types of sampling:

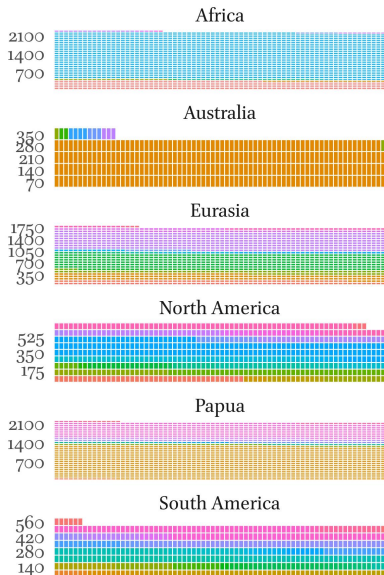


### Stratified random sampling

divide the population into groups that differ in important ways, and then perform random sampling for each group



## There are different types of sampling:

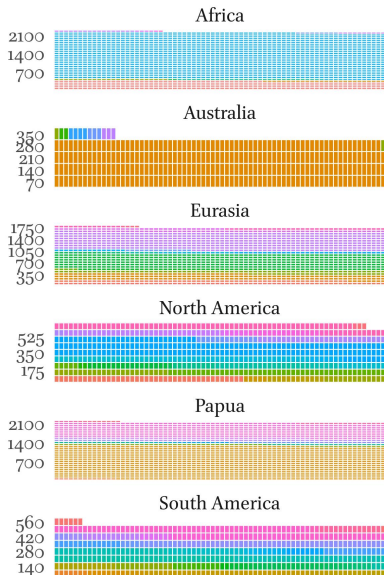


### Stratified random sampling

divide the population into groups that differ in important ways, and then perform random sampling for each group

- !!! The Glottolog version in the `lingtypology` package suggests that there are **214 unique combinations** (142 sign languages and 82 isolates counted as one family)

## There are different types of sampling:



### Stratified random sampling

divide the population into groups that differ in important ways, and then perform random sampling for each group

!!! The Glottolog version in the `lingtypology` package suggests that there are **214 unique combinations** (142 sign languages and 82 isolates counted as one family)

⇒ So to create a statistically reasonable sample one needs to get around 300 languages

## I am not the first to discuss this problem

- [Bell 1978] "Language Samples"
- [Dryer 1989] "Large Linguistic Areas and Language Sampling"
- [Perkins 1989] "Statistical Techniques for Determining Language Sample Size"
- [Nichols 1992] "Linguistic Diversity in Space and Time"
- [Rietveld and Van Hout 1993] "Statistical Techniques for the Study of Language and Language Behaviour"
- [Rijkhoff and Bakker 1998] "Language sampling"
- [Maslova 2000] "A dynamic approach to the verification of distributional universals"
- [Widmann 2001] "Language Sampling for Typological Studies"
- [Janssen et al. 2006] "Randomization Tests in Language Typology"
- [Baker 2010] "Language Sampling"

## Sampling bias

- Geneological
- Caused by contact
- Cultural
- Typological
- Populational

## Sampling bias

- Geneological
- Caused by contact
- Cultural
- Typological
- Populational
- Bibliographical

## Sampling bias

- Geneological
- Caused by contact
- Cultural
- Typological
- Populational
- Bibliographical
- **Typological** — only typologists think that one typological value corresponds to one so called language

## Theoretical linguists

- Complain about how hard it is to solve a problem
- Don't publish any results until it will be ideal

## Computational linguists

- Solve the wrong problem
- Publish messy data and messy results

## Theoretical linguists

- Complain about how hard it is to solve a problem
- Don't publish any results until it will be ideal

## Computational linguists

- Solve the wrong problem
- Publish messy data and messy results

## My suggestion:

- Don't do any sampling
- Use a linguistic family (or analogous units) as a minimal unit of typological research
- Analyse all languages in a family
- Publish your data
- Make a call for contributions
- Update your results



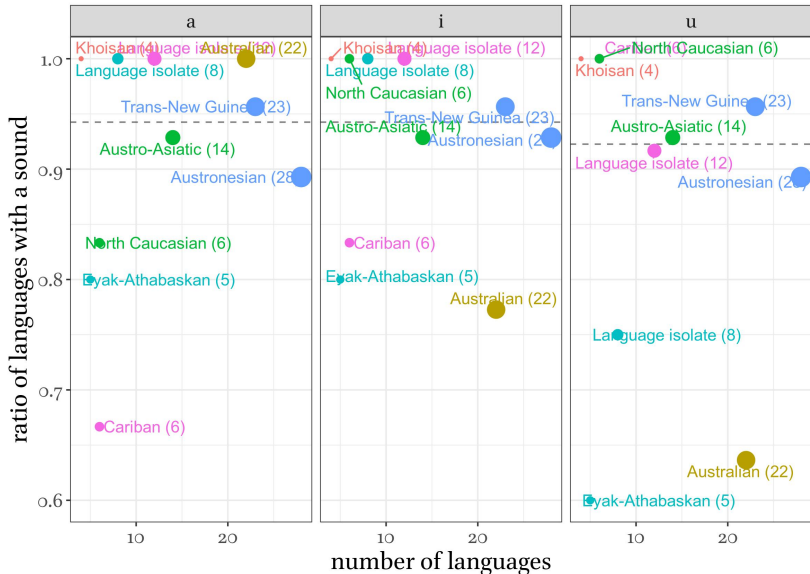
## frequentist view

- There is a population with one fixed value  $\theta$
- Sample from the population and estimate the value  $\hat{\theta}$
- If you want to replicate the previous study, resample the data and reestimate the value  $\hat{\theta}$

## Bayesian view

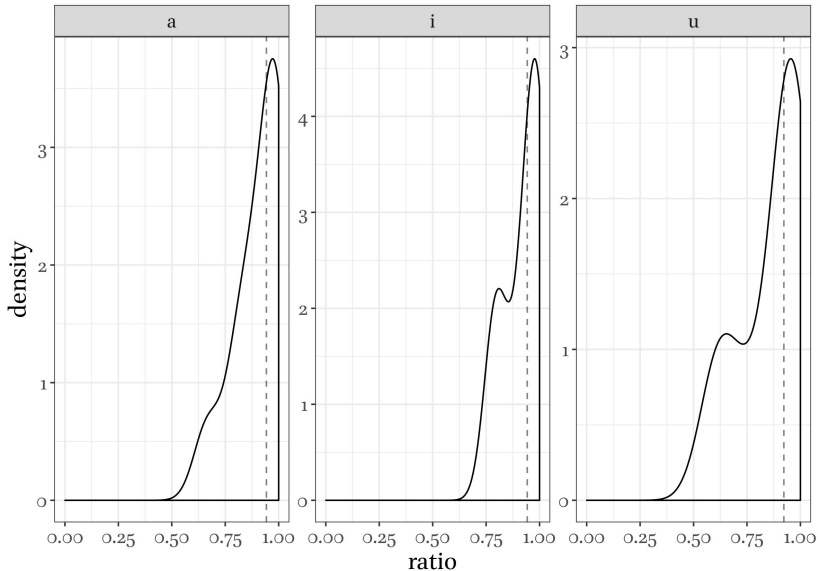
- There is a value  $\theta$  that could be described as a distribution of probabilities
- Take into account previous works and formulate **prior** knowledge about  $\theta$
- Sample from the population and estimate the value  $\theta$
- Use Bayes' formula to get **posterior** distribution of  $\theta$
- Use an obtained result as a future prior and update your previous data

# Case study: how frequent are *a*, *i* and *u*? (10 families)



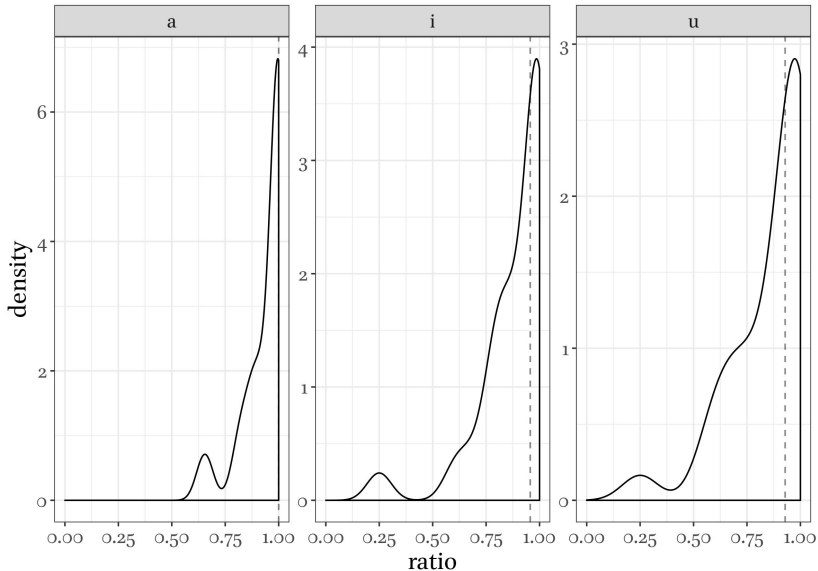
sample of 128 languages

## Case study: how frequent are *a*, *i* and *u*? (10 families)



sample of 128 languages

## Case study: how frequent are *a*, *i* and *u*? (29 families)



sample of 354 languages

## What about phonology?

It is possible to use phonological units or relations from any phonological theory you like:

- Features, feet, syllables, etc.
- Feature constituents, OT constraints, exemplars, phonological are diachronic alternations
- Phonological distinctions (e. g. /i/ vs. /i/)
- ...

Send me a letter!  
[agricolamzgmail.com](mailto:agricolamzgmail.com)

Presentation is available here:  
[tinyurl.com/y3wtkcbq](https://tinyurl.com/y3wtkcbq)



## References

- Baker, D. (2010). Language sampling. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology*. Oxford University Press.
- Bell, A. (1978). Language samples. In J. H. Greenberg, C. A. Ferguson, and E. A. Moravcsik (Eds.), *Universals of human language, vol. 4: Syntax*. Stanford University Press.
- Dryer, M. S. (1989). Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 13(2), 257–292.
- Janssen, D. P., B. Bickel, and F. Zúñiga (2006). Randomization tests in language typology. *Linguistic Typology*, 419–40.
- Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4(3), 307–333.
- Nichols, J. (1992). *Linguistic diversity in space and time*. University of Chicago Press.
- Perkins, R. D. (1989). Statistical techniques for determining language sample size. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 13(2), 293–315.
- Rietveld, T. and R. Van Hout (1993). *Statistical techniques for the study of language and language behaviour*. Walter de Gruyter.
- Rijkhoff, J. and D. Bakker (1998). Language sampling. *Linguistic typology* 2(3), 263–314.
- Widmann, T. M. (2001). Language sampling for typological studies. Master's thesis, University of Aarhus.