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

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 θ 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 θ 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 θ 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 θ 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 θ 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 θ 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 θ that you know, add some
  confidence intervals of you need to convince an editor who is mad
  about statistics

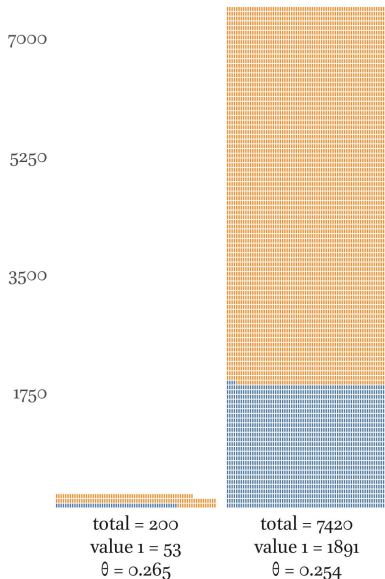- After you published your paper project is finished

There are different type of sampling

7000

5250

3500

1750

total = 200
value 1 = 53
θ = 0.265

total = 7420
value 1 = 1891
θ = 0.254
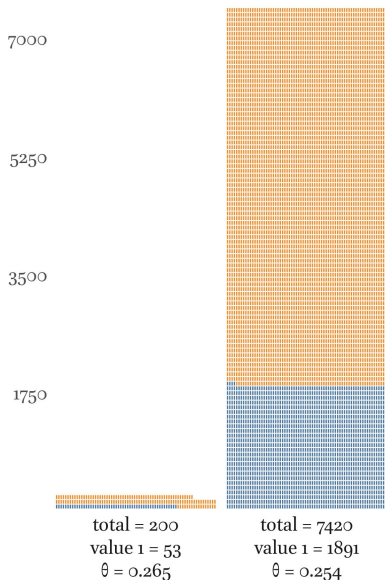
Random sampling
each member of the population has
an equal probability of selection
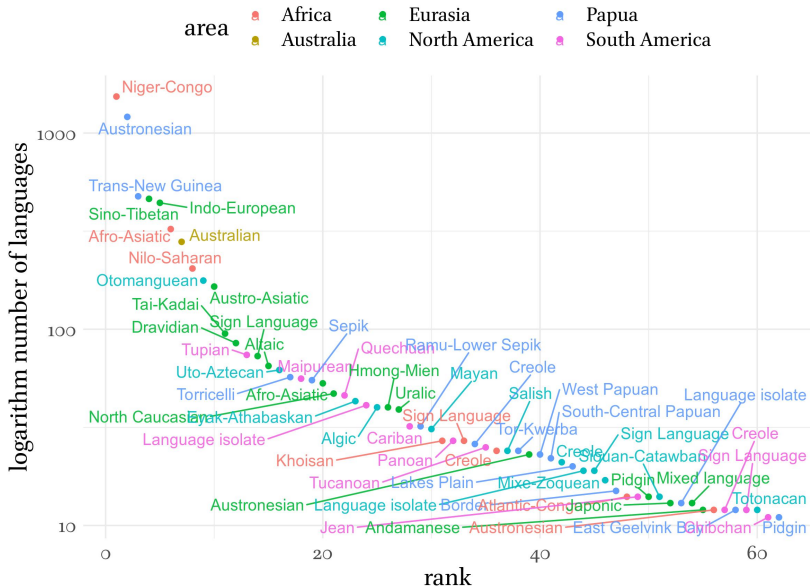
There are different type of sampling

Random sampling
each member of the population has
an equal probability of selection

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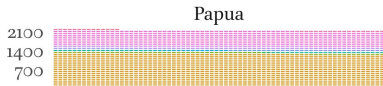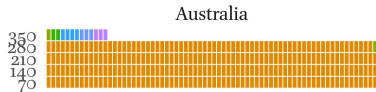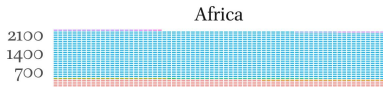area**, so observations are not
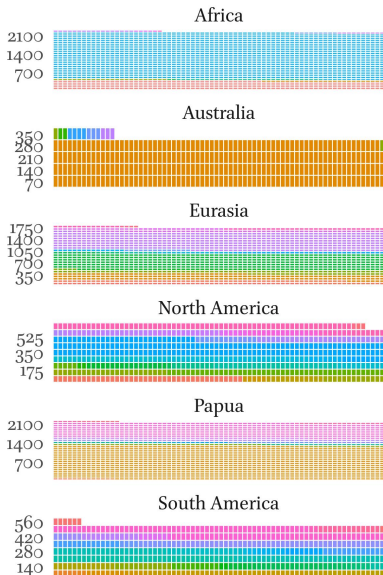independent...

Language families (languages > 10)

area

Africa · Eurasia · Papua
Australia · North America · South America

6

There are different type 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 type of sampling:

Africa
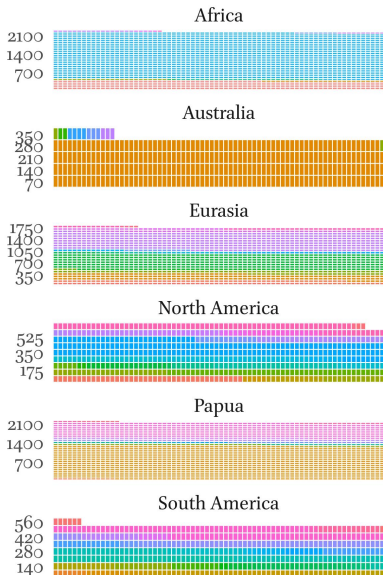Australia
Eurasia
North America
Papua
South America

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 type of sampling:

Africa

Australia

Eurasia

North America

Papua

South America

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"

Biases of sampling

- geneological
- caused by contacts
- cultural
- typological
- populational

Biases of sampling

- geneological
- caused by contacts
- cultural
- typological
- populational
- bibliographical

Biases of sampling

- geneological
- caused by contacts
- cultural
- typological
- populational
- bibliographical

- typologistical — because only typologists think that there is only one typological value per one so called language

"theoretical linguists"

- complain about how hard is to solve a problem
- don't publish any results, until it will be ideal

"computational linguists"

- solve the wrong problem
- publish a messy results

"theoretical linguists"

- complain about how hard is to solve a problem
- don't publish any results, until it will be ideal

"computational linguists"

- solve the wrong problem
- publish a messy results

my suggestion:

- don't do any sampling
- use a linguistic families (or analogues unites) as a minimal unit of typological research
- analyse all languages in a family
- publish your data
- make a call for a contribution
- update your results after the contribution

frequentist view

- there is a population with a one fixed value θ
- sample from the population and estimate value $\hat{\theta}$
- if you want to replicate previous study, resample data and reestimate value $\hat{\theta}$
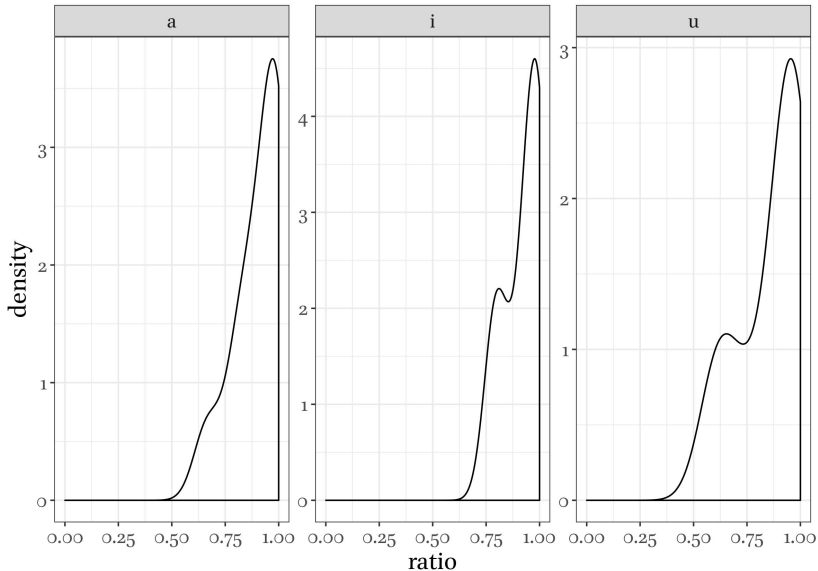
bayesian view

- there is a value θ that could be described as a destribution of probabilities
- take into account previous works and formulate **prior** knowledge abut θ
- sample from the population and estimate value θ
- use Bayes formula to get **posterior** distribution of θ
- use an obtained result as a future prior, and as a result update your previous data

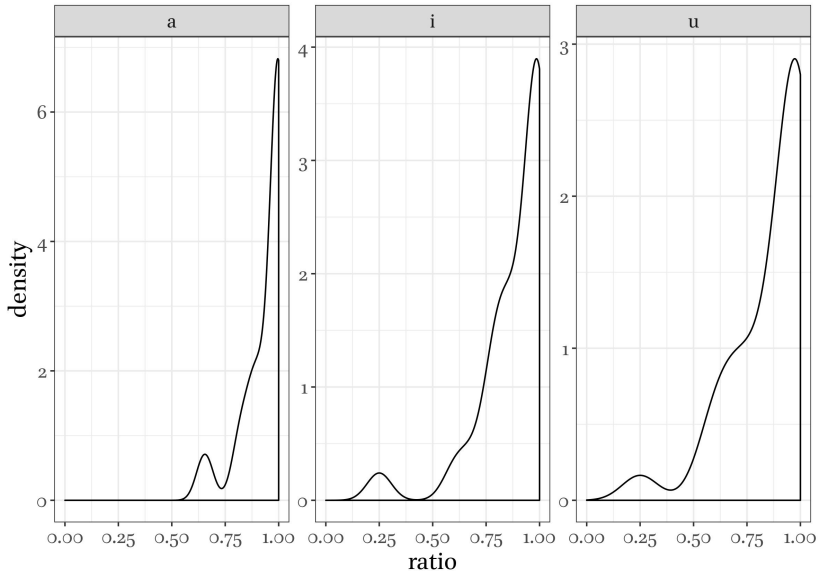# Case study: how friquent are *a, i* and *u*? (10 families)
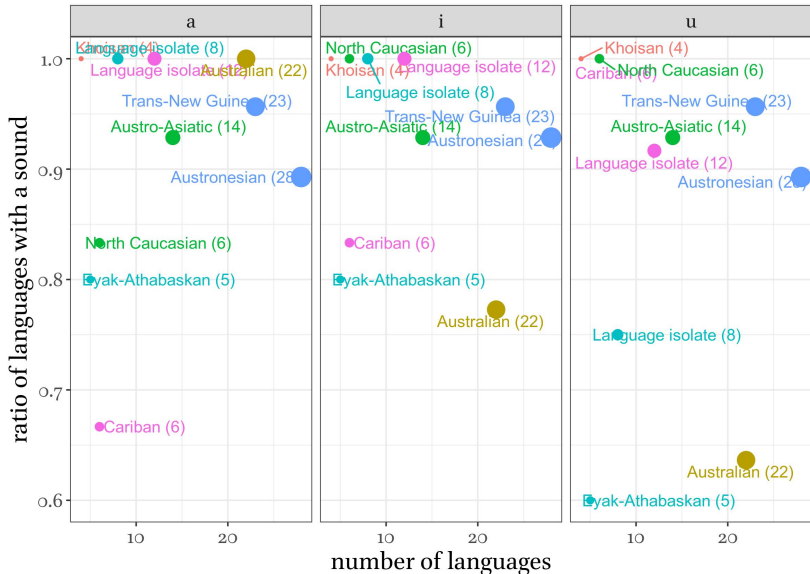
sample of 128 languages

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

sample of 128 languages

13

Case study: how friquent 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. /ɨ/)
- ...

Send me a letter!
agricolamzgmail.com

Presentation is available here:
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.