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 the linguistic typology
- different strategies of sampling
- Bayesian way of thinking about the linguistic typology
- Case study: vowels

Goals of the linguistic typology

- attest distributions (statistical and areal) of typological values

Goals of the linguistic typology

- attest distributions (statistical and areal) of typological values
- find a correlation of distributions of different typological categories
  - absolute universals
  - distributional patterns and tendencies
  - semantic maps
  - diachronic change of typological values

Goals of the linguistic typology

- attest distributions (statistical and areal) of typological values
- find a correlation of distributions of different typological categories
  - absolute universals
  - distributional patterns and tendencies
  - semantic maps
  - diachronic change of typological values
- find a correlation of linguistic and non-linguistic patterns
  - population movements
  - population size
  - language contact
  - sociolinguistic parameters
  - geopolitical environment (including diseases' spread)

Goals of the linguistic typology

- attest distributions (statistical and areal) of typological values
- find a correlation of distributions of different typological categories
  - absolute universals
  - distributional patterns and tendencies
  - semantic maps
  - diachronic change of typological values
- find a correlation of linguistic and non-linguistic patterns
  - population movements
  - population size
  - language contact
  - sociolinguistic parameters
  - geopolitical environment (including diseases' spread)
- deal with a mixed typological value per language

Frequentist typological research

- Formulate a theoretical question
  There is a category in some languages with values **VAL 1** and **VAL 2**.

Frequentist typological research

- Formulate a theoretical question
  There is a category in some languages with values **VAL 1** and **VAL 2**.
  What is the probability θ to meet **VAL 1** in randomly picked language?

Frequentist typological research

- Formulate a theoretical question
  There is a category in some languages with values **VAL 1** and **VAL 2**. What is the probability θ to meet **VAL 1** in randomly picked language?
- Get a grant, hire some students, or select a holiday, which you want to spent working on this topic...

Frequentist typological research

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

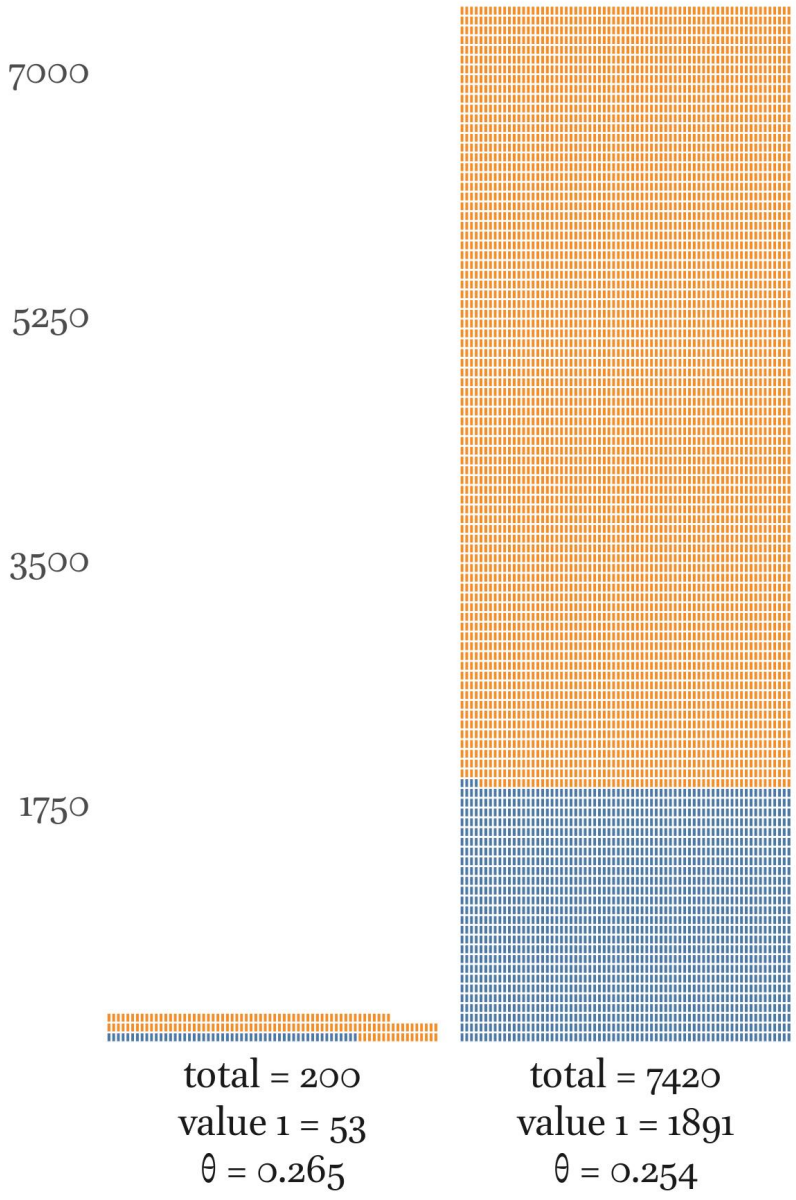
Frequentist typological research

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

Frequentist typological research

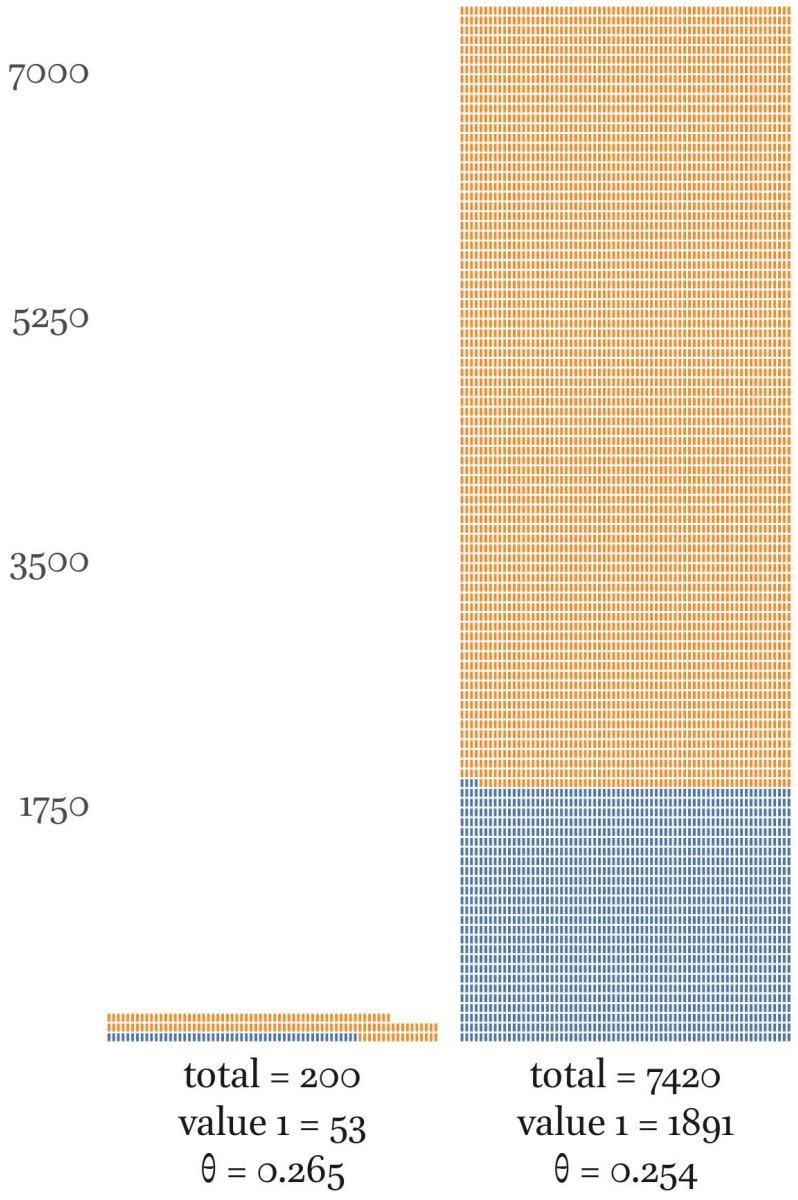
- Formulate a theoretical question
  There is a category in some languages with values **VAL 1** and **VAL 2**. What is the probability θ to meet **VAL 1** in randomly picked language?
  - Get a grant, hire some students, or select a holiday, which you want to spent working on this topic...
- Pick a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

There are different type of sampling



Random sampling:
each element in the population has an equal probability of selection

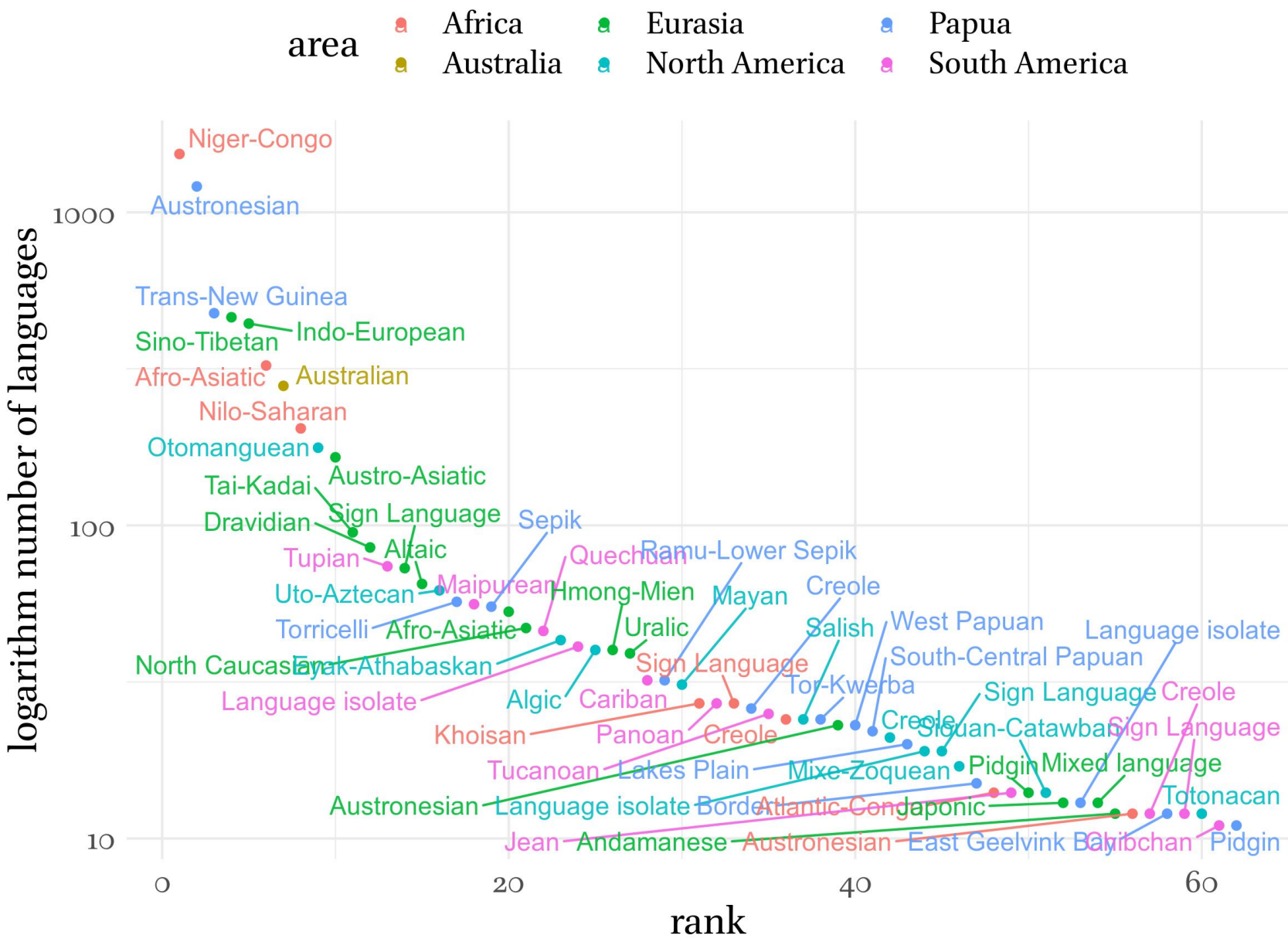
There are different type of sampling



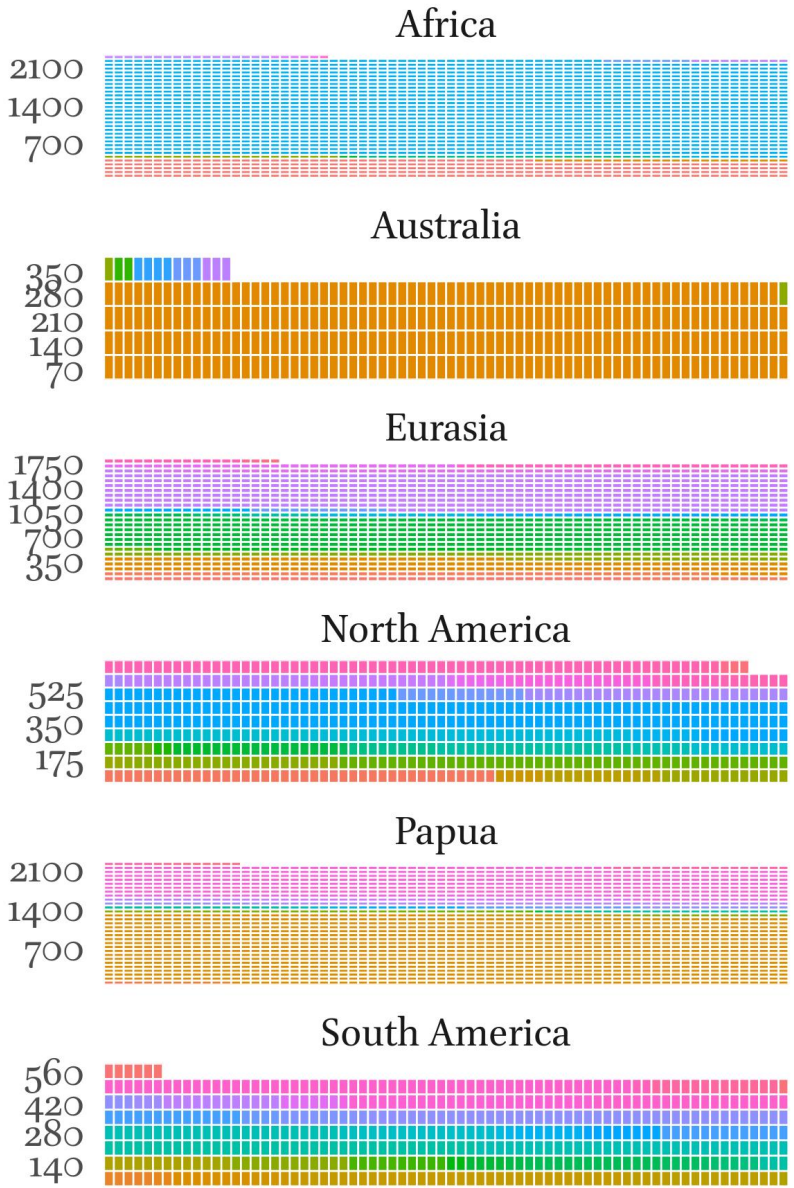
Random sampling:
each element in the population has an equal probability of selection

!!! but each language is grouped in **a language family** and **an area**, so each observation is not independent…

Language families (languages > 10)

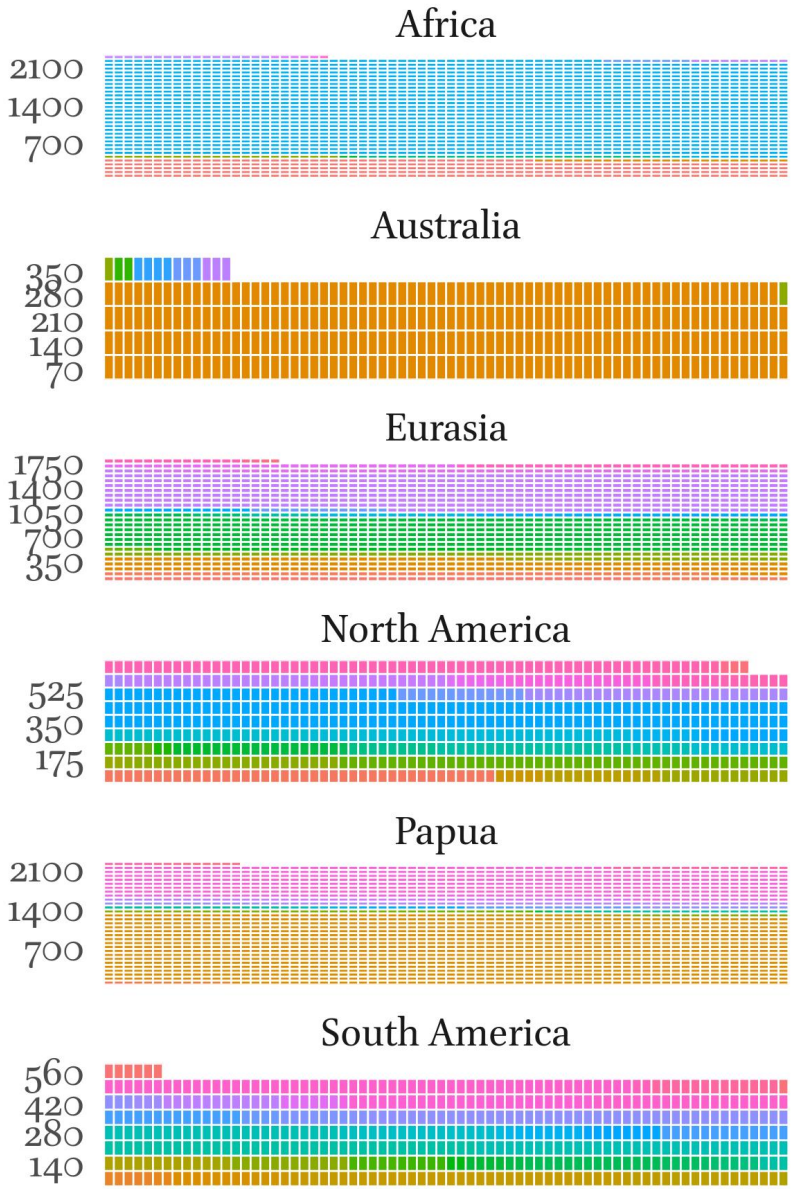


There are different type of sampling:



Stratified random sampling divide population into groups that differ in important ways and then perform random sampling from each group

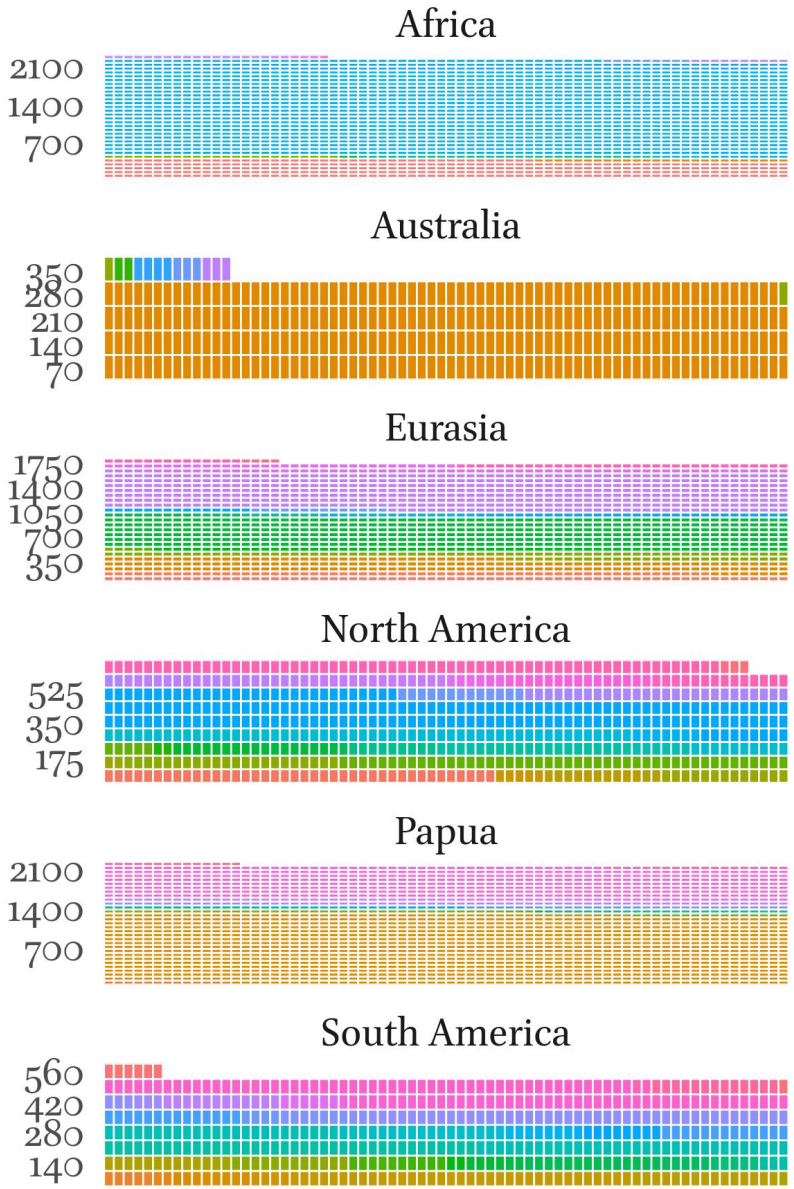
There are different type of sampling:



Stratified random sampling divide population into groups that differ in important ways and then perform random sampling from 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:



Stratified random sampling divide population into groups that differ in important ways and then perform random sampling from 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 for creating statistically reasonable sample one need to get around 300 languages

I am not the first

- [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"

What about phonology?

It is possible to use phonological units or relation of any phonological theory you like:

- features, feet, syllables, etc.
- feature constituents, OT constraints, exemplars, phonological 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.