Revisiting the population *vs* phoneme-inventory correlation

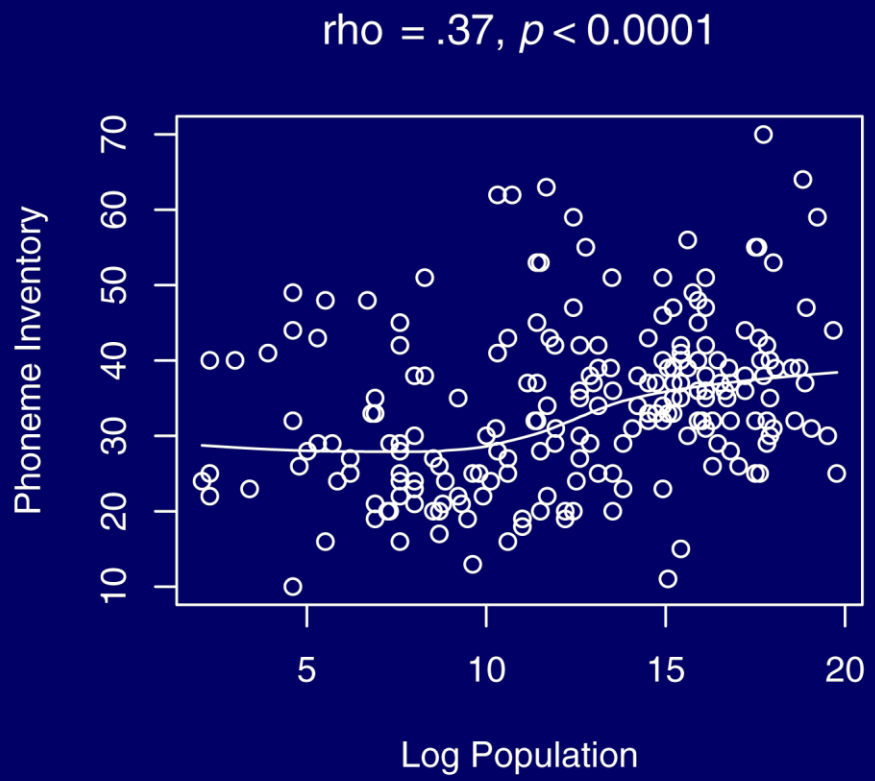
Steven Moran • Daniel McCloy • Richard Wright

University of Washington, Seattle

Overview

- Review of two previous studies
  - Hay & Bauer (2007)
  - Donohue & Nichols (2011)
- Our study
  - Methods
  - Results
  - Interpretation
- Concluding remarks

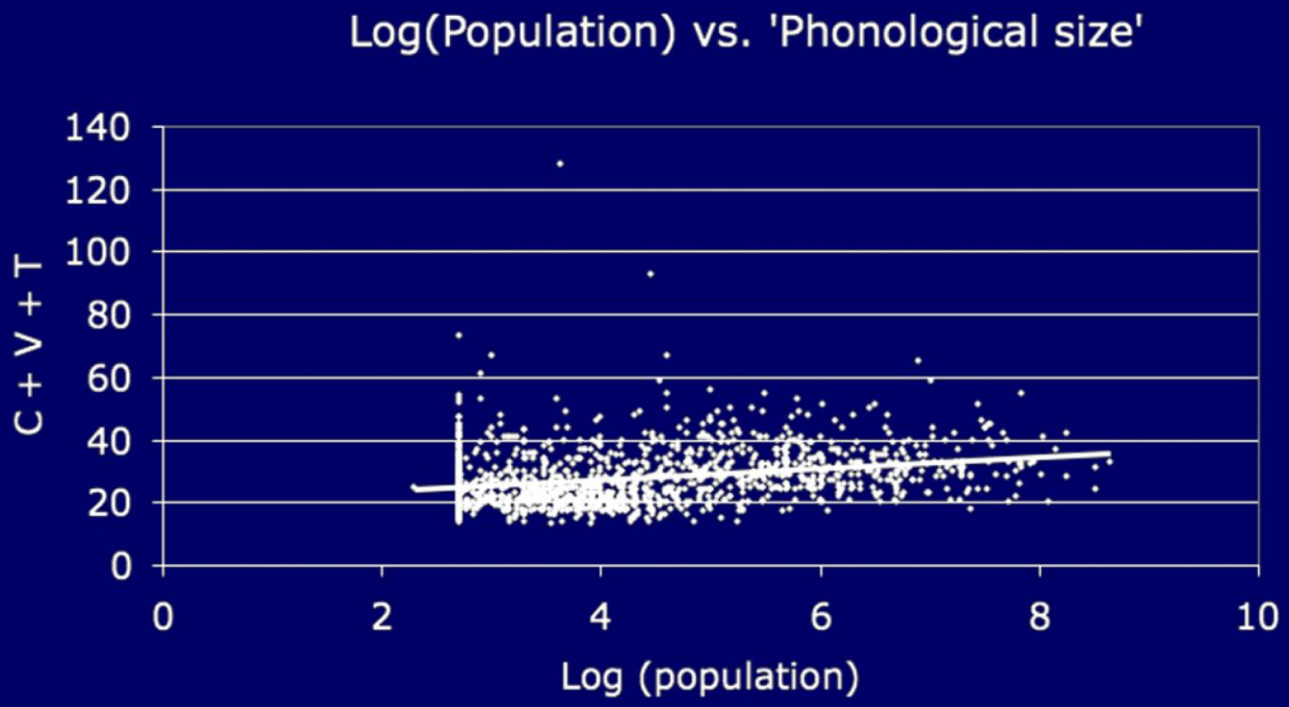
Hay & Bauer (2007)



- $N = 216$ languages
- $\rho = 0.37$ (statistically significant)

Graph reprinted from Hay & Bauer (2007)

Donohue & Nichols (2011)



- $N = 1350$ languages
- $r = 0.27$ (not significant)

Graph reprinted from Donohue & Nichols (2011)

Which one is correct?

Hay & Bauer (2007)

- Sample
  - 216 language "convenience sample" from Bauer (2007)
  - Major world languages, well-known isolates, & typologically interesting languages
- Analysis
  - Spearman rank correlations
  - Data not independent (languages "nested" within families)

Donohue & Nichols (2011)

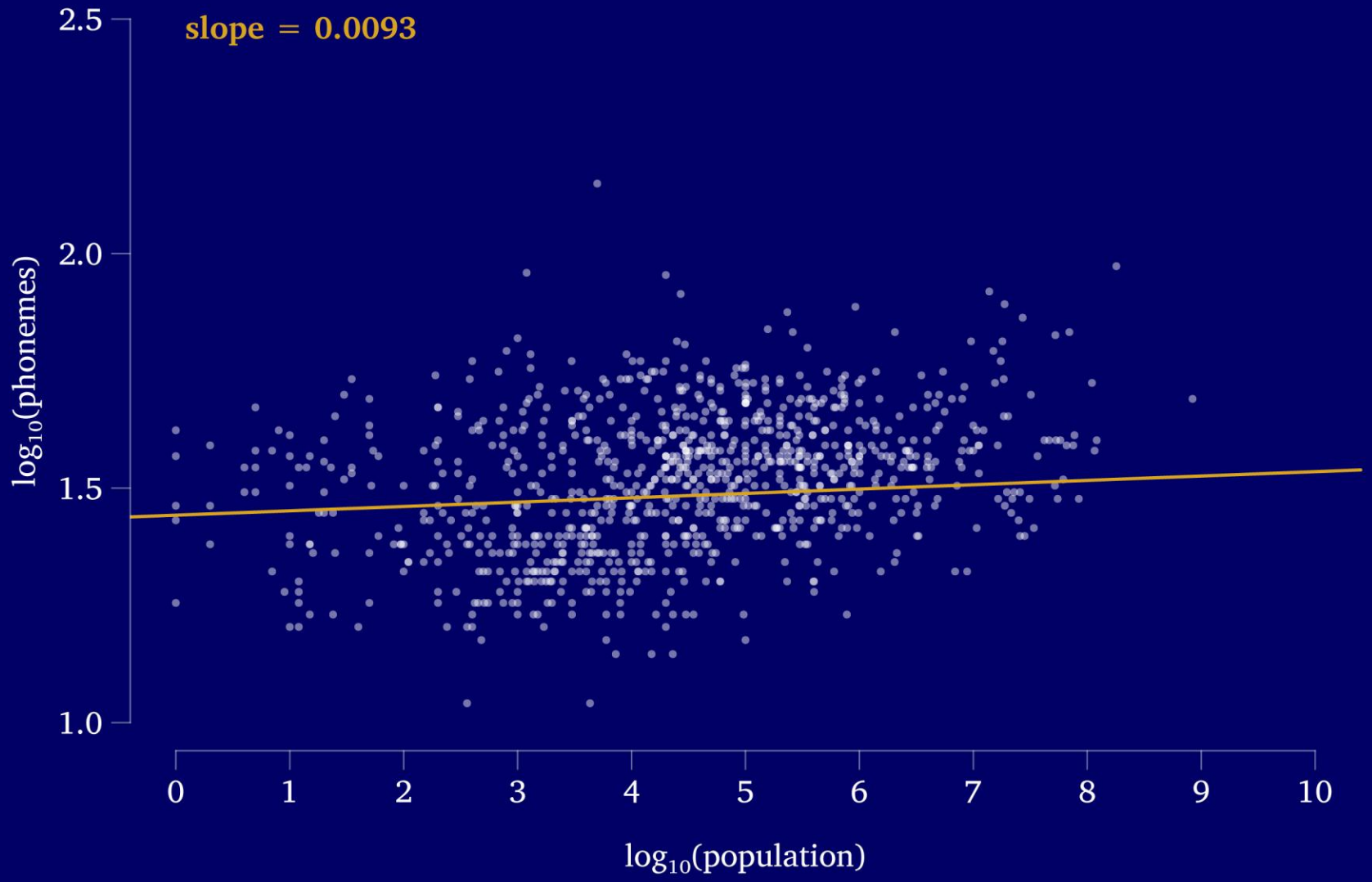
- Sample
  - 1350 languages, well-distributed both genealogically and areally (based on AutoTyp)
- Analysis
  - Simple linear regressions
  - Data not independent (languages "nested" within families)

Our study

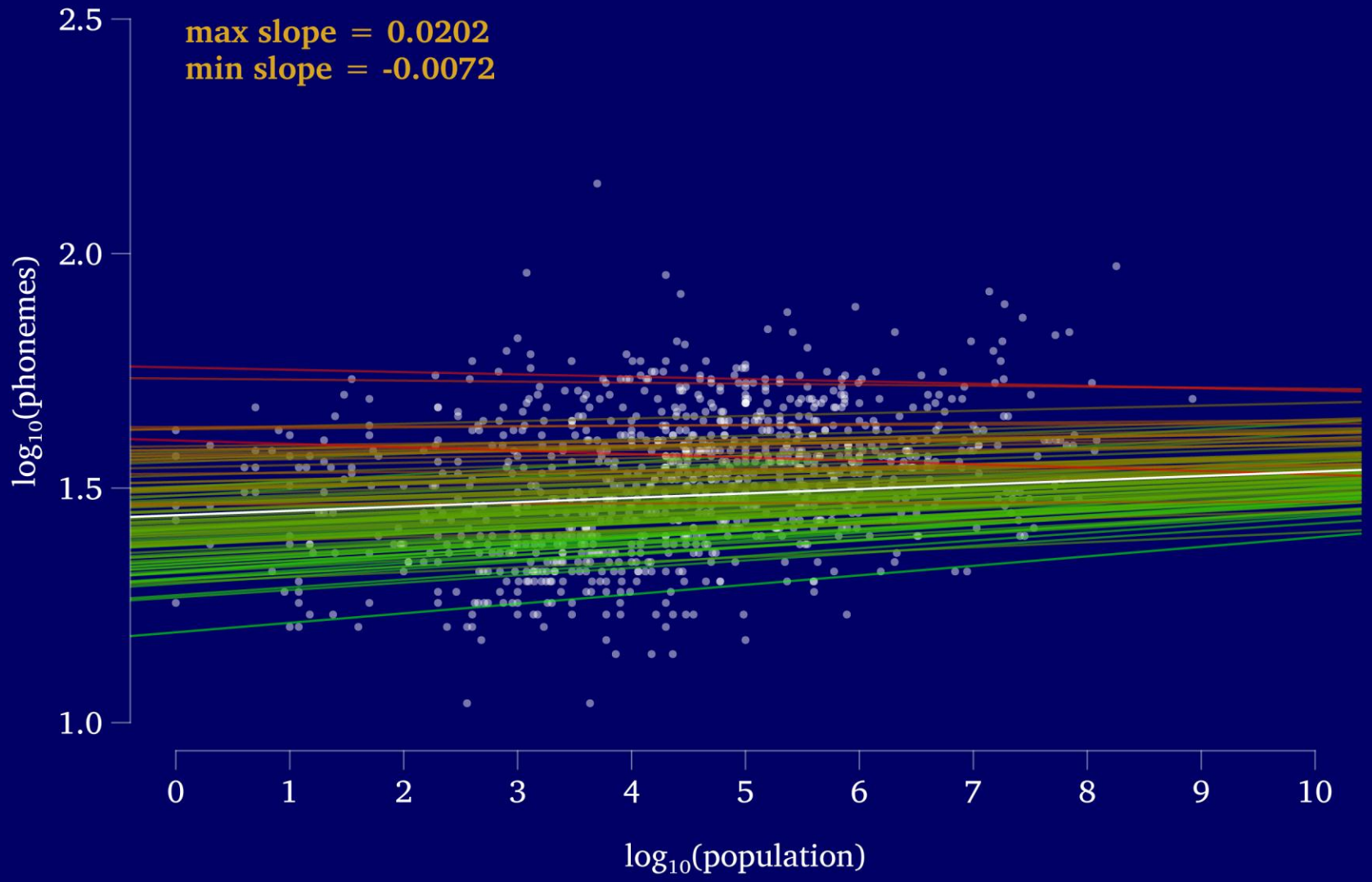
- Sample
  - 969 languages from the PHOIBLE knowledge base[1]
  - Subsumes *Alphabets des langues africaines,*[2] SPA[3] & UPSID[4]
  - 100 families, 321 genera, 18 isolates
  - Excludes extinct, ancient, mixed, pidgin, and creole languages
- Analysis
  - Heirarchical mixed effects model
    - Accomodates non-independent (nested) data
    - Models the within- and between-group variance

[1] Moran & Wright (2009)
[2] Hartell (1993), Chanard (2006)
[3] Crothers et al. (1979)
[4] Maddieson (1984), Maddieson & Precoda (1990)

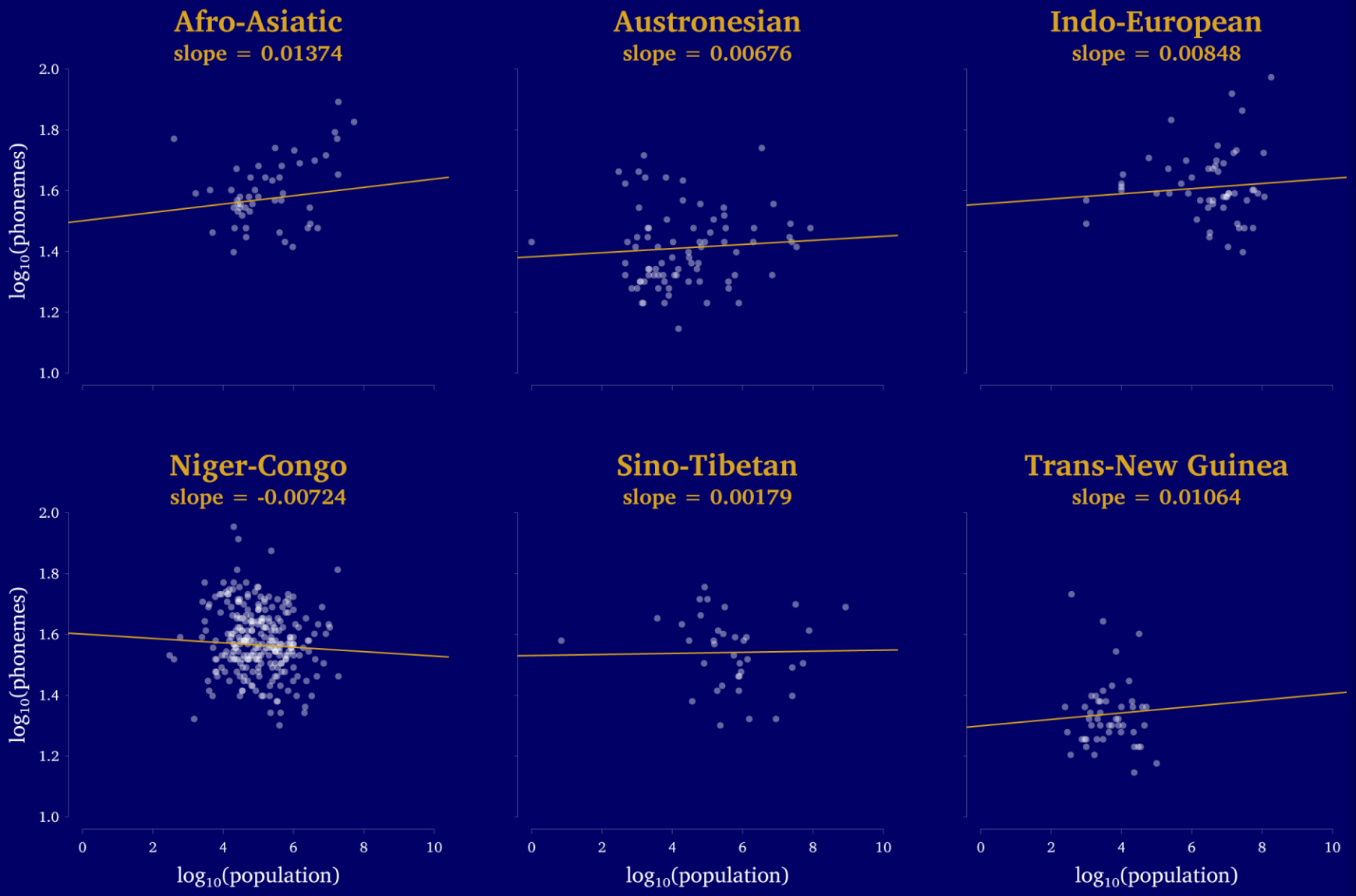
Overall regression



Individual family regressions



Regressions for the six largest families



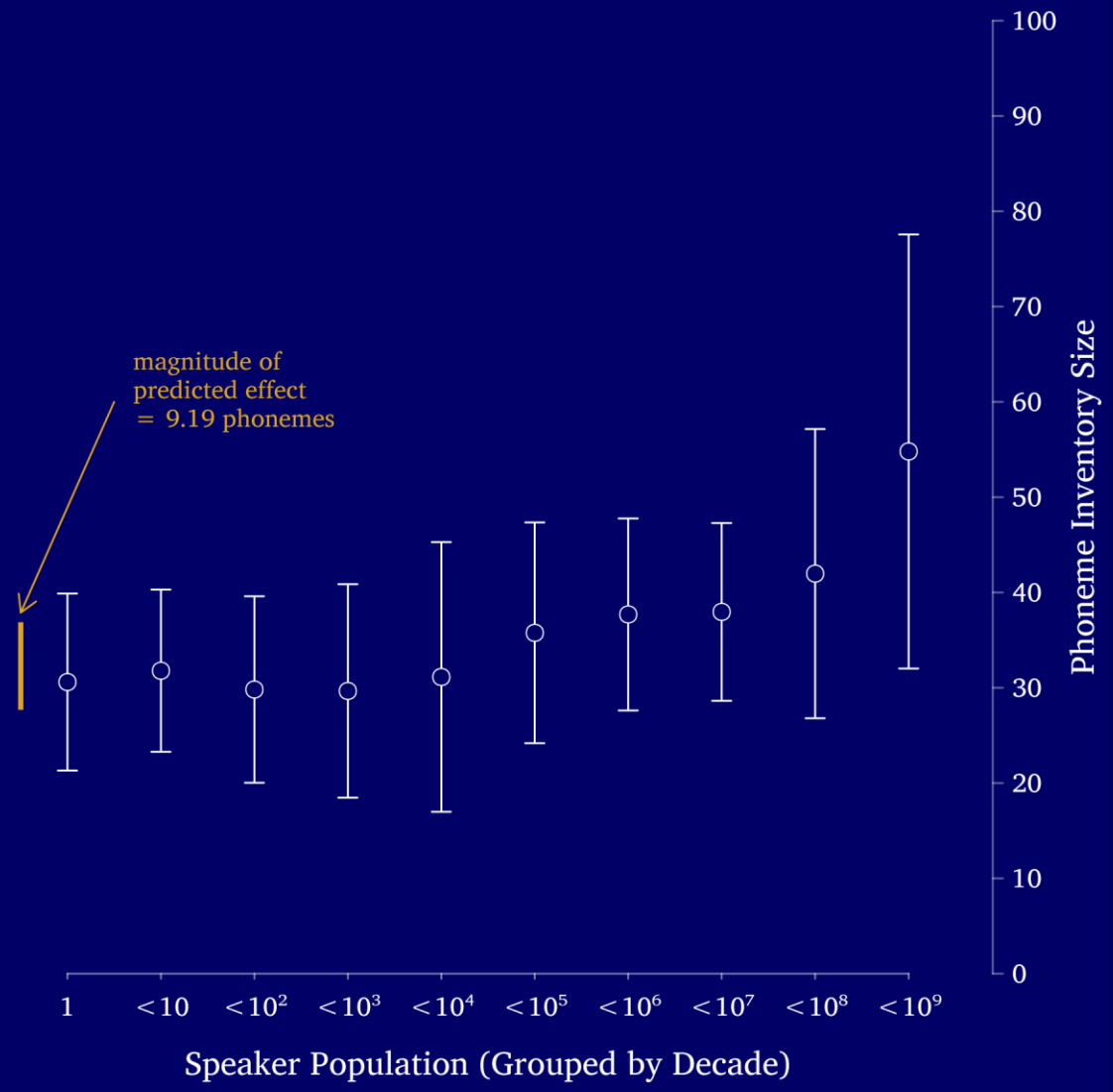
Model summary

Fixed effect estimate (left) and variance estimates (center, right) for model predicting phoneme inventory size ($N = 969$)

	Fixed effect		Random effect for genus ($n = 321$)			Random effect for family ($n = 100$)		
Predictor	Coefficient (S.E.)	t	s^2	s	corr.	s^2	s	corr.
intercept	1.4423 (0.0204)	70.8403	0.0000	0.0000	0.0000	0.0162	0.1272	-0.6540
log(pop.)	0.0093 (0.0041)	2.2632	0.0001	0.0088		0.0001	0.0111	

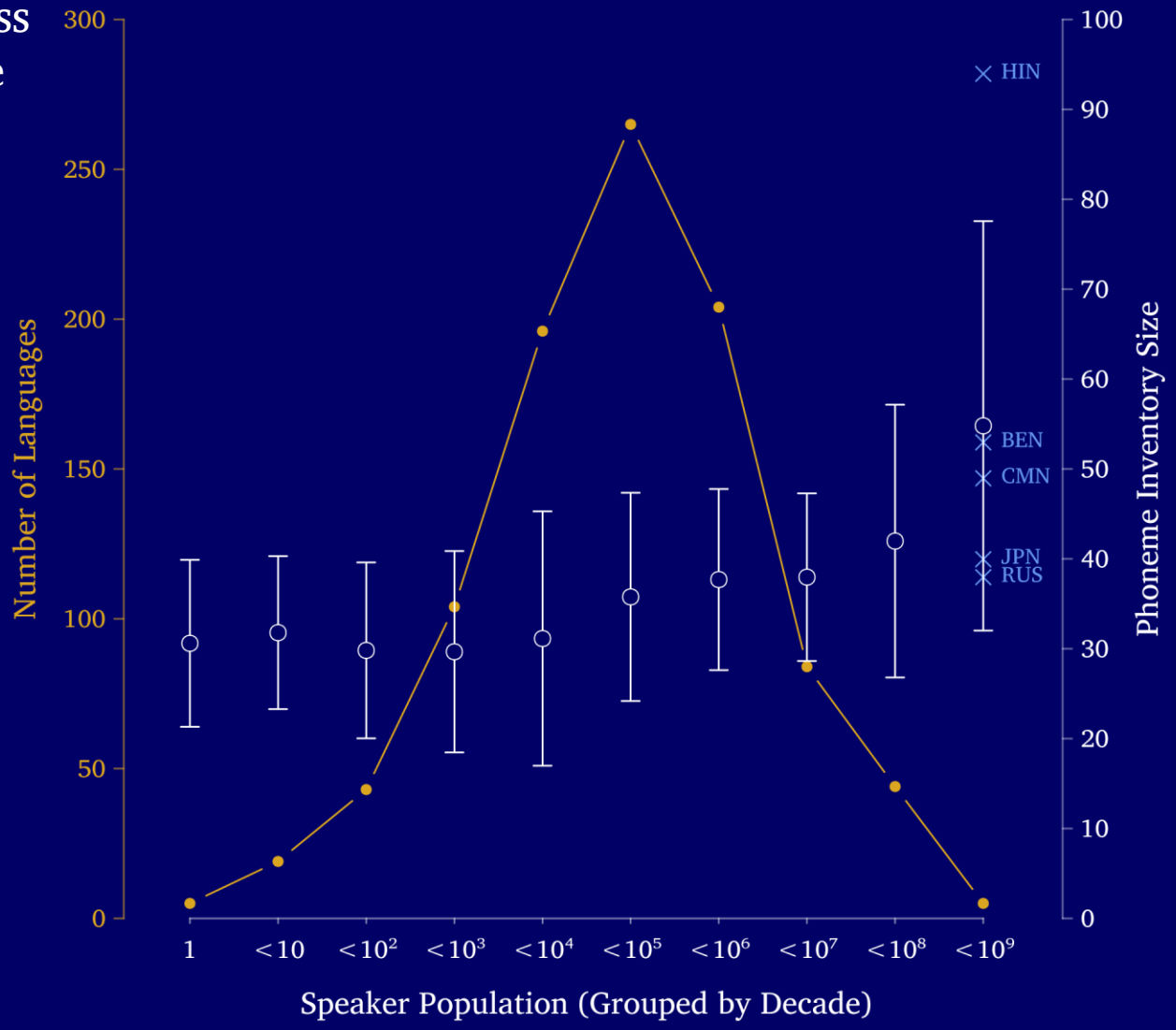
Magnitude of predicted effect

- Predicted effect across full population range is less than the standard deviation within any given population-based cohort



Magnitude of predicted effect

- Predicted effect across full population range is less than the standard deviation within any given population-based cohort
- 10^8–10^9 cohort skewed upward by outlier (HIN: Hindi)



Interpreting our results

- The relationship is most likely a statistical artefact
  - Evidence: the within-family trends range from increasing, through flat, to decreasing

- Even if it's not an artefact, the relationship is too small to be meaningfully interpreted
  - Evidence: size of predicted effect (1.02 phonemes per order-of-magnitude) is much smaller than the variability within similar-population-size language cohorts

The bigger picture

- Why expect a correlation at all?[1]
  - Population can change rapidly (war, disease, migration…)
  - Mechanism for phonological change often absent
- If population isn't a good predictor, then what is?
  - A complex web of factors likely influence phoneme inventory size[2]
    - Language family
    - Language contact situation
    - Social network structure
    - etc.

[1] see Donohue & Nichols (2011) for additional critiques of the logic of this correlation
[2] see Trudgill (2011) for an overview of his research in this area

Concluding remarks

"We know that for large enough sample sizes, every study — including ones in which the null hypothesis of no effect is true — will declare a statistically significant effect." [1]

[1] van der Laan & Rose (2010).

Acknowledgments & Thanks to:

- The University of Washington's Royalty Research Fund for partial funding of PHOIBLE development
- PHOIBLE development assistance from Morgana Davids, Scott Drellishak, David Ellison, Richard John Harvey, Kelley Kilanski, Michael McAulife, Kevin Pittman, Brandon Plasters, Cameron Rule, Daniel Smith, and Daniel Veja
- Marilyn Vihman for providing the Stanford Phonology Archive data
- Tristan Purvis and Christopher Green for assistance with African languages
- Paul Sampson, Theresa Smith, and Donghun Kim for statistical consultation

References

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, *332*(6027), 346-349. doi:10.1126/science.1199295

Bauer, L. (2007). *The linguistics student's handbook*. Oxford: Oxford University Press.

Chanard, C. (2006). *Systèmes alphabétiques des langues africaines.* Retrieved from http://sumale.vjf.cnrs.fr/phono/

Crothers, J. H., Lorentz, J. P., Sherman, D. A., & Vihman, M. M. (1979). *Handbook of phonological data from a sample of the world's languages: A report of the Stanford Phonology Archive*. Palo Alto, CA: Department of Linguistics, Stanford University.

Donohue, M., & Nichols, J. (2011). Does phoneme inventory size correlate with population size? *Linguistic Typology*, *15*, 161-170. doi:10.1515/LITY.2011.011

Hartell, R. L. (1993). *Alphabets des langues africaines.* Dakar, SN: UNESCO, Bureau Régional de Dakar.

Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, *83*(2), 388-400.

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (16th ed.). Dallas, TX: SIL International. Retrieved from http://www.ethnologue.com/

Maddieson, I. (1984). *Patterns of sounds.* Cambridge, UK: Cambridge University Press.

Maddieson, I., & Precoda, K. (1990). Updating UPSID. *UCLA Working Papers in Phonetics*, *74*, 104-111.

Moran, S., & Wright, R. (2009). *Phonetics information base and lexicon (PHOIBLE)*. Seattle, WA. Retrieved from http://phoible.org/

Trudgill, P. (2011). Social structure and phoneme inventories. *Linguistic Typology*, *15*, 155-160. doi:10.1515/LITY.2011.010

van der Laan, M., & Rose, S. (2010). Statistics ready for a revolution. *Amstat News.* Retrieved from http://magazine.amstat.org/blog/2010/09/01/statrevolution/

Backup Slides

About the PHOIBLE knowledge base

- Currently over 1500 languages (and growing!)
- Each language record includes:
  - Phonemes: all segments in unicode IPA; some records also include allophones & tonemes
  - Features: each phoneme as a vector of distinctive features, structured as an extensible mathematical graph
  - Genealogy: Language name, ISO 639-3 code, family codes from Multitree,[1] genus-level classifications from WALS[2]
  - Provenance: PDF snapshots from source grammars
  - Demographics: Speaker population, lat./long., GDP, etc.

[1] Multitree: A digital library of language relationships. (2009). Ypsilanti, MI: Institute for Language Information and Technology (LINGUIST List), Eastern Michigan University. Retrieved from http://multitree.org/

[2] Dryer, M. S., & Haspelmath, M. (Eds.). (2011). The world atlas of language structures online. Munich: Max Planck Digital Library. Retrieved from http://wals.info/