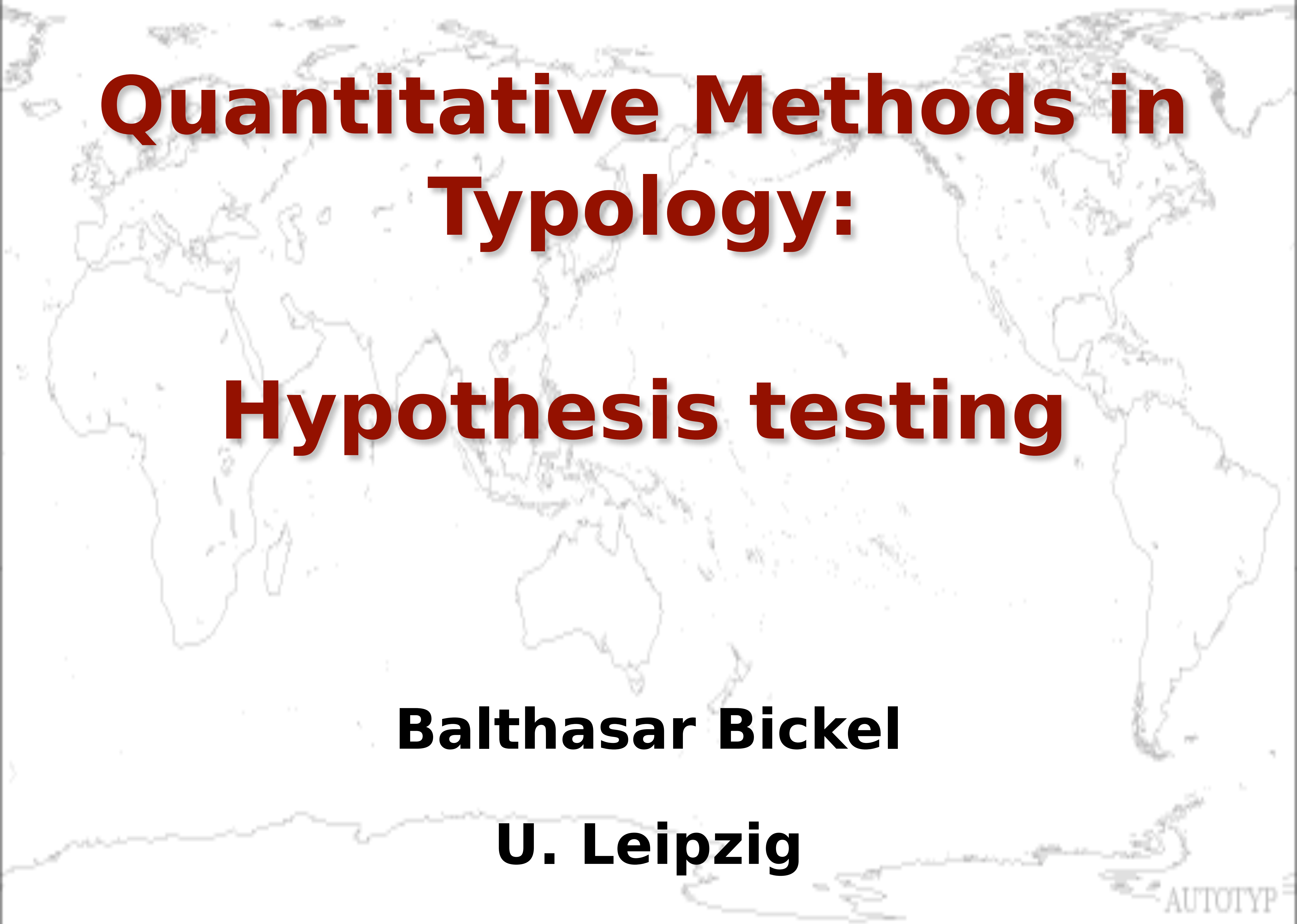


Quantitative Methods in Typology:

Hypothesis testing

Balthasar Bickel

U. Leipzig

A key question of quantitative typology

- Typology, and especially Multivariate Typology results in many hypotheses on empirical universals (a.k.a. probabilistic or statistical universals), such as
  - local and disjunct ILL-scope tends to block WH words in dependent clauses

- How can we establish that universals of this kind are **non-accidental properties of our sample**?

Testing statistical universals: two key problems

1. The **INFERENCE** problem: how to extrapolate from a sample to all human languages in the absence of random sampling?

   NB: there already have been at least 500,000 languages, but we know only about 500!

Testing statistical universals: two key problems

2. The **DIACHRONY** problem: how to tell apart the factors determing the observed cross-linguistic distribution?

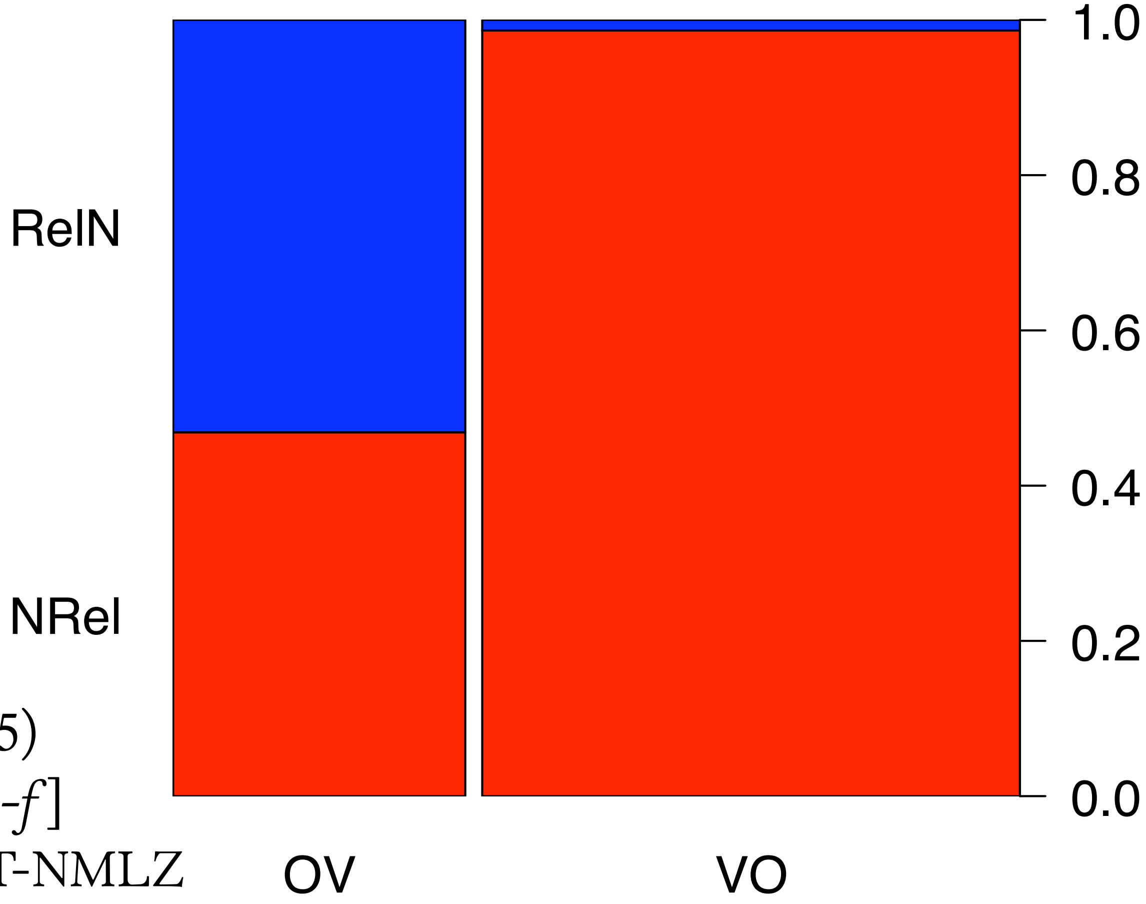
- Structural Pressure (‘what works best, replicates best’; a.k.a. ‘selection’, ‘functional pressure’, ‘preferred pathways of change’, ‘linguistic principles’)
- Language Contact (‘what is most popular, replicates best’)
- Blind Inheritance (‘what was best for our parents, is best for us’)

Testing statistical universals: two key problems

NB: don’t underestimate the possibility of Blind Inheritance, especially in *large* samples. Example:

Nepali
[*us-le* *sāikal calā-eko*] *mānche-lāi* *cin-yo.*
3s-ERG bike drive-PART person-DAT know-3sPST
‘He knew the person who rode the bike.’

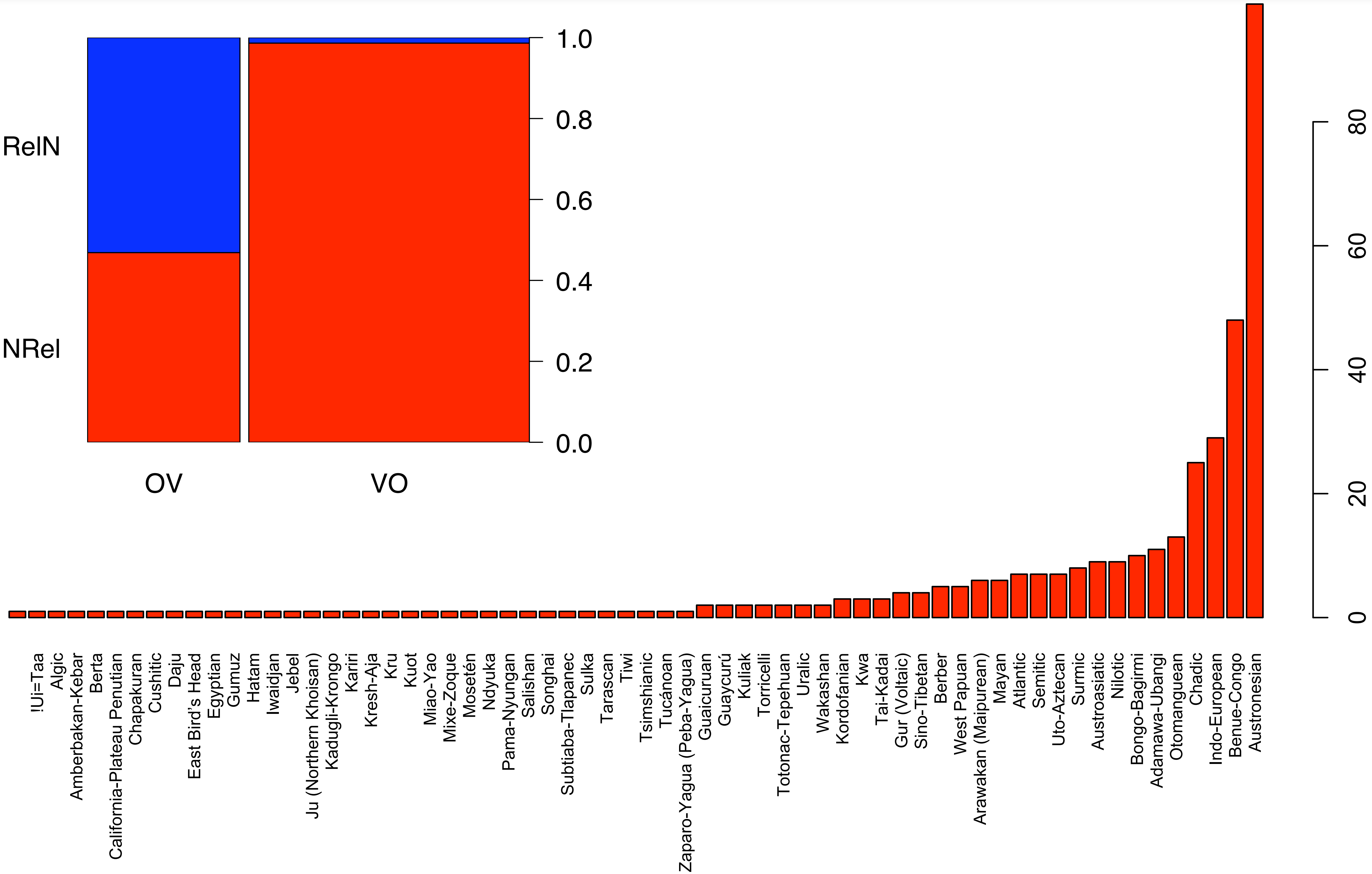
Mandarin Chinese
ta renshi [*qi* *che* *de* *na-ge*] *ren*
3s know ride bike ATTR DEM-CL person
‘He knew the person who rode the bike.’



Harar Oromo (Owens 1985)
dubar-tíi [*xennáa xann-ée-f*]
girl-NOM present give-PST-NMLZ
‘the girl he gave a present to’

N = 580 (Data from Dryer 2005, *WALS*)

Testing statistical universals: two key problems



Toward a solution

- If we can solve the Diachrony Problem, we will have solved the Inference Problem:
  - if we know that a distribution is due to Structural Pressure as a factor in diachrony, *independent of Language Contact and Blind Inheritance,* Structural Pressure determines how languages develop over time, beyond what we can observe synchronically.
  - Note that it does not matter at this point whether Structural Pressure works as a synchronic “filter” on diachronic change or channels change itself — the key issue is first to make sure the observed distribution is caused by Structural Pressure and not by Contact or Blind Inheritance,
  - and that it is not accidental, i.e. not due to random fluctuation in diachrony.

Solving the Diachrony Problem

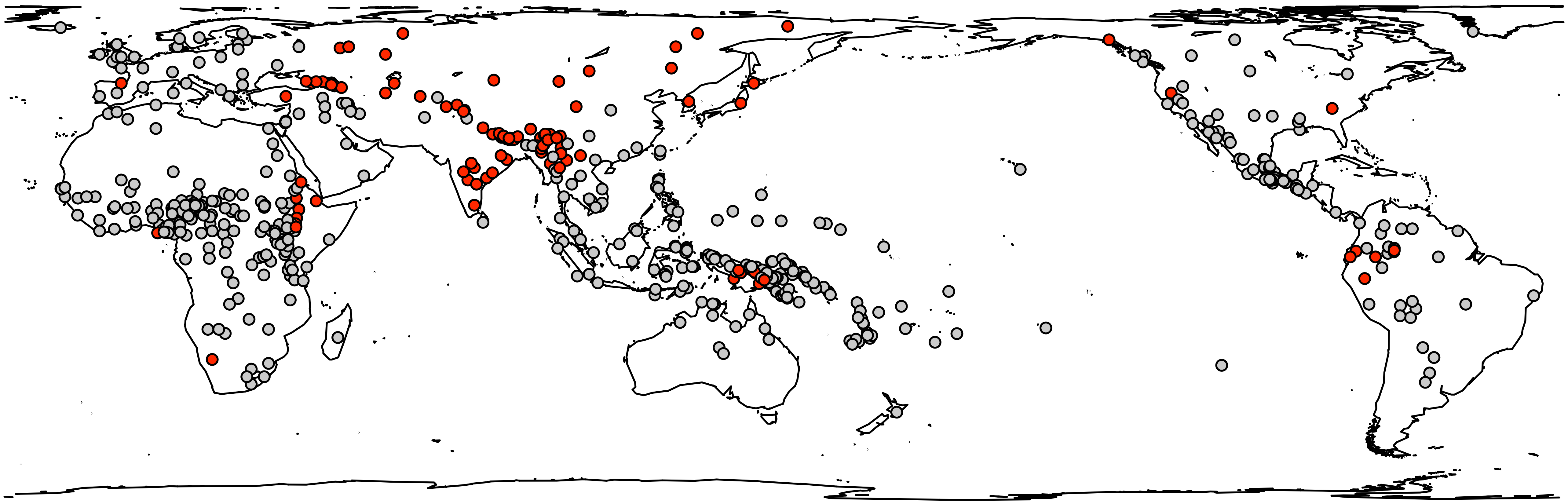
- Useful reformulation of empirical universals: instead of $p \rightarrow q$,
- $E(\mathrm{q}) \sim \mathrm{p}$,
  - where E represent the expected mean of q, given p
  - $E(\mathrm{q})$ can be directly linked to p if q is continuous; categorical q need to be transformed first by what is called a ‘link function’ g(), usually the natural logarithm of odds (‘logit’, ‘logistic regression’) or counts (‘loglinear analysis’). This defines the Generalized Linear Model:

    $$\mathrm{g}(\mathrm{E}(q)) = \alpha + \beta_1 p_1 + \beta_2 p_2 \ldots + \beta_k p_k$$

  - unrestricted universal: $\mathrm{g}(\mathrm{E}(q)) = \alpha$, and α deviates from H_0.

Using Generalized Linear Models for testing universals

- If *p* is categorical, β represents the difference in E(*q*) between *p*=0 (e.g. OV) and *p*=1 (e.g. VO). All multinomial *p* with *k* levels can be ‘parametrized’ into *k*-1 paramaters (a.k.a. ‘dummy variables’)
- Note that *p* can be a complex vector of predictors (‘competing motivations’), including also linguistic areas, e.g.
  - the odds for OV&RelN are higher in Eurasia than elsewhere

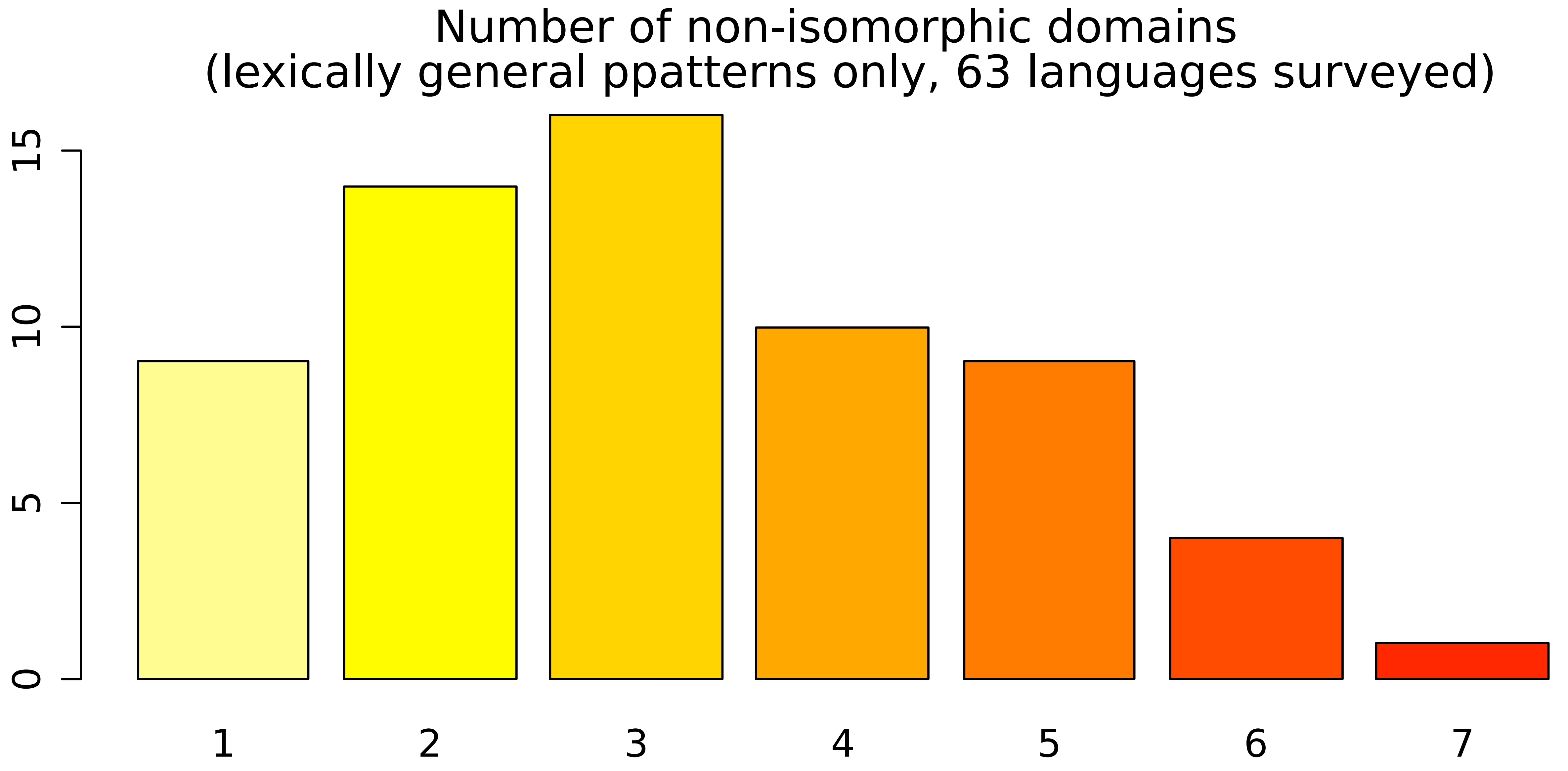


Using Generalized Linear Models for testing universals

- And it can include interactions $p_i \cdot p_j$, i.e. differences in effects between levels of p_i across levels of p_j.
- The goal of statistical analysis is
  - to find those coefficients $\beta_1 \dots \beta_k$ that best describe the data (e.g. via Least Squares or Maximum Likelihood Estimation) and then
  - to test whether these coefficients are significantly different from zero, i.e. ‘belong to the model’ statistically,
    - by searching for the most parsimoneous model that still fits the data well.

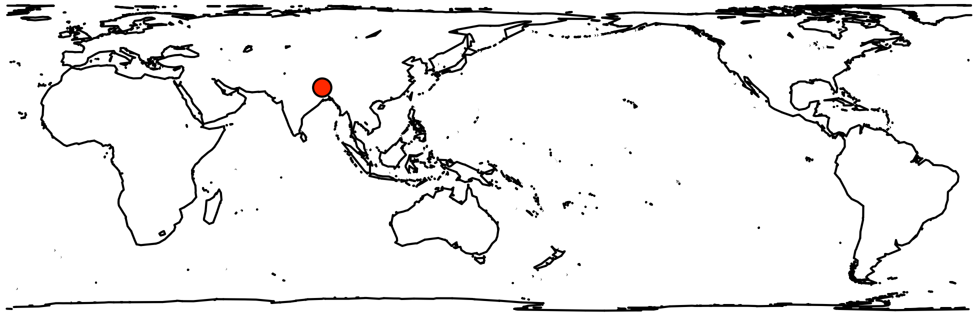
An example: relative p-word size

- Case study on phonological words
- Found that many languages have more than one phonological word (Schiering, Bickel & Hildebrandt 2010, J of Ling.):



An example: relative p-word size

- Limbu (Sino-Tibetan, Nepal):



  - pf-[stem-sf=cl], domain of Liquid Alternation and other rules

    kɛ-[**L**i’-**L**e=**L**o] > kɛ[**l**i’**r**e**r**o] ‘of your bow’

    2sPOSS-bow-GEN=PTCL

  - [pf-stem-sf=cl], domain of Coronal Assimilation and other rules

    [mɛ-**n**-mɛ**t**-paŋ] > [mɛ**m**mɛ**p**paŋ] ‘We did not tell him’

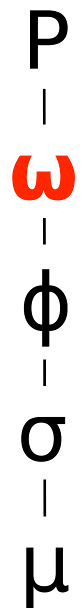
    nsA-NEG-tell-1>3.PST

- What’s the word in Limbu? The red one or the blue one? Does the word in Limbu include or exclude the prefix?

Bickel, Hildebrandt, & Schiering 2009

An example: relative p-word size

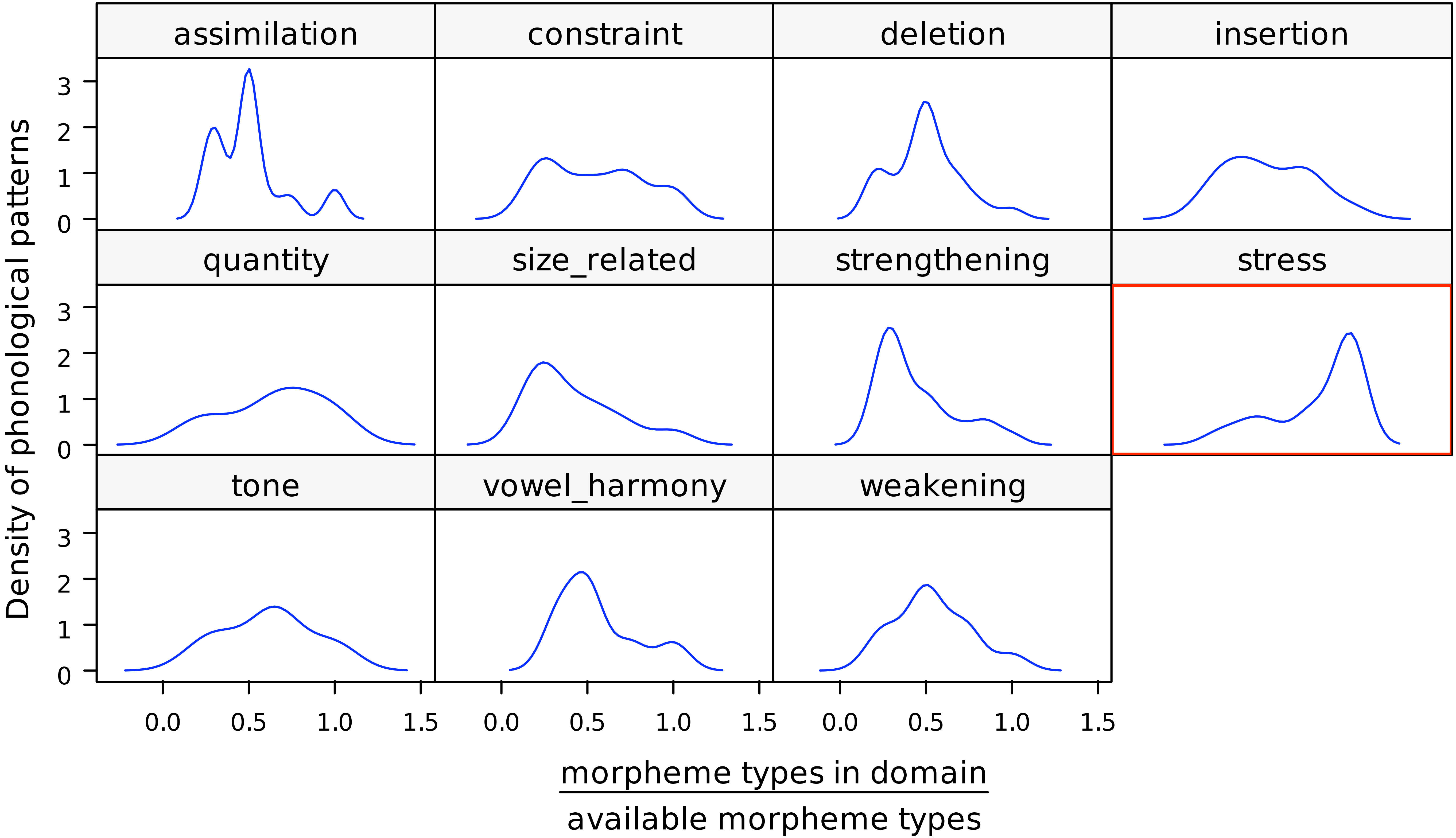
- Challenge for the theory of the Prosodic Hierarchy, which states — as an absolute universal — that a language has exactly one p-word (modulo recursion; e.g. Vogel 2009):



- But we approach this empirically: what’s where why?
- Compute relative coherence of p-words:

$$c(\mathrm{p};\ \mathrm{L}) = \frac{N\,(\text{morpheme types referenced by } p)}{N\,(\text{morpheme types in } L)}$$

An example: relative p-word size



Bickel, Hildebrandt, & Schiering 2009

An example: relative p-word size

→*Hypothesis:* since stress patterns systematically interact with higher-level domains (phrases, utterances; information structure), they target larger domains than what is targeted by other phonological patterns.

- *Hypothesized to be very common:*

  Limbu (Sino-Tibetan) Stress: [prefix-'stem-suffix=clitic]

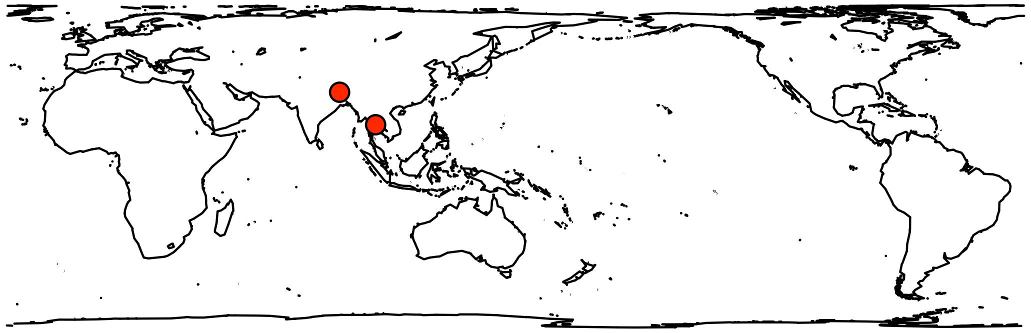
  [mɛ-'thaŋ-e=aŋ]
  3ns-come.up-PST=and

- *Hypothesized to be much less common:*

  Mon (Austroasiatic) Stress: ['cl]=[pf<infix>'stem]=['cl]

  [k<ə>'lɒʔ]
  <CAUS>cross

  ['kɒ]=['klɒʔ]
  CAUS=cross



Bickel, Hildebrandt, & Schiering 2009

An example: relative p-word size

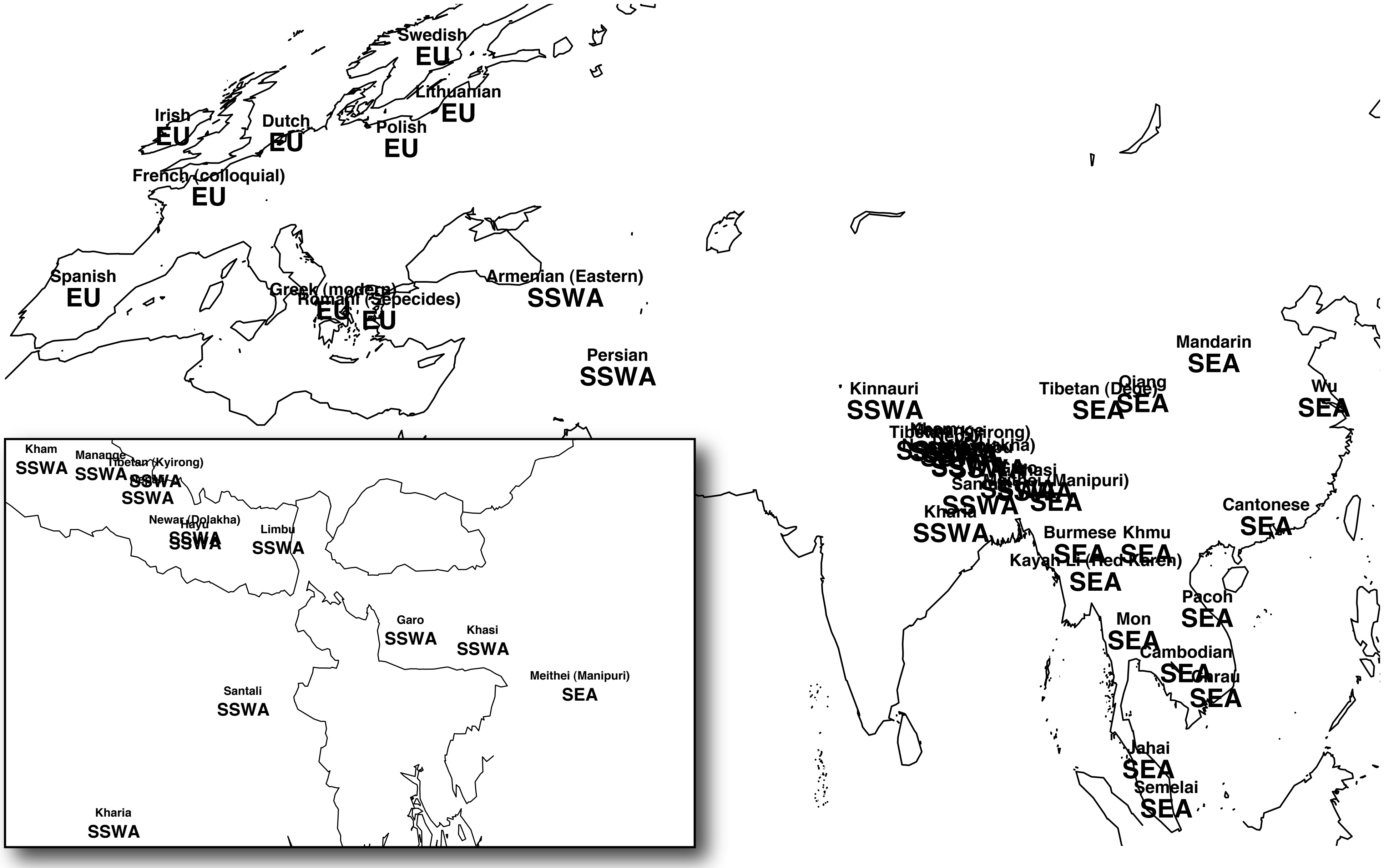
- Apart from Structural Pressure, ie. [±stress pattern], other plausible predictors are:
    - Language Contact *within areas*, e.g. [±Southeast Asia]
    - Blind Inheritance *within proven families, i.e. groups with a single ancestor*, e.g. [±Austroasiatic]
- Model to test: $E(\mathrm{c}) = \mu(\mathrm{c})$, thus:

$$\mu(\mathrm{c}) = \alpha + \beta_1[\mathrm{FAMILIES}] + \beta_2[\mathrm{AREAS}] + \beta_3[\mathrm{STRESS}] + \beta_4[\mathrm{FAMILIES}][\mathrm{STRESS}] + \beta_5[\mathrm{AREAS}][\mathrm{STRESS}]$$

- Random Fluctuation = what is left unexplained by the equation

An example: relative p-word size

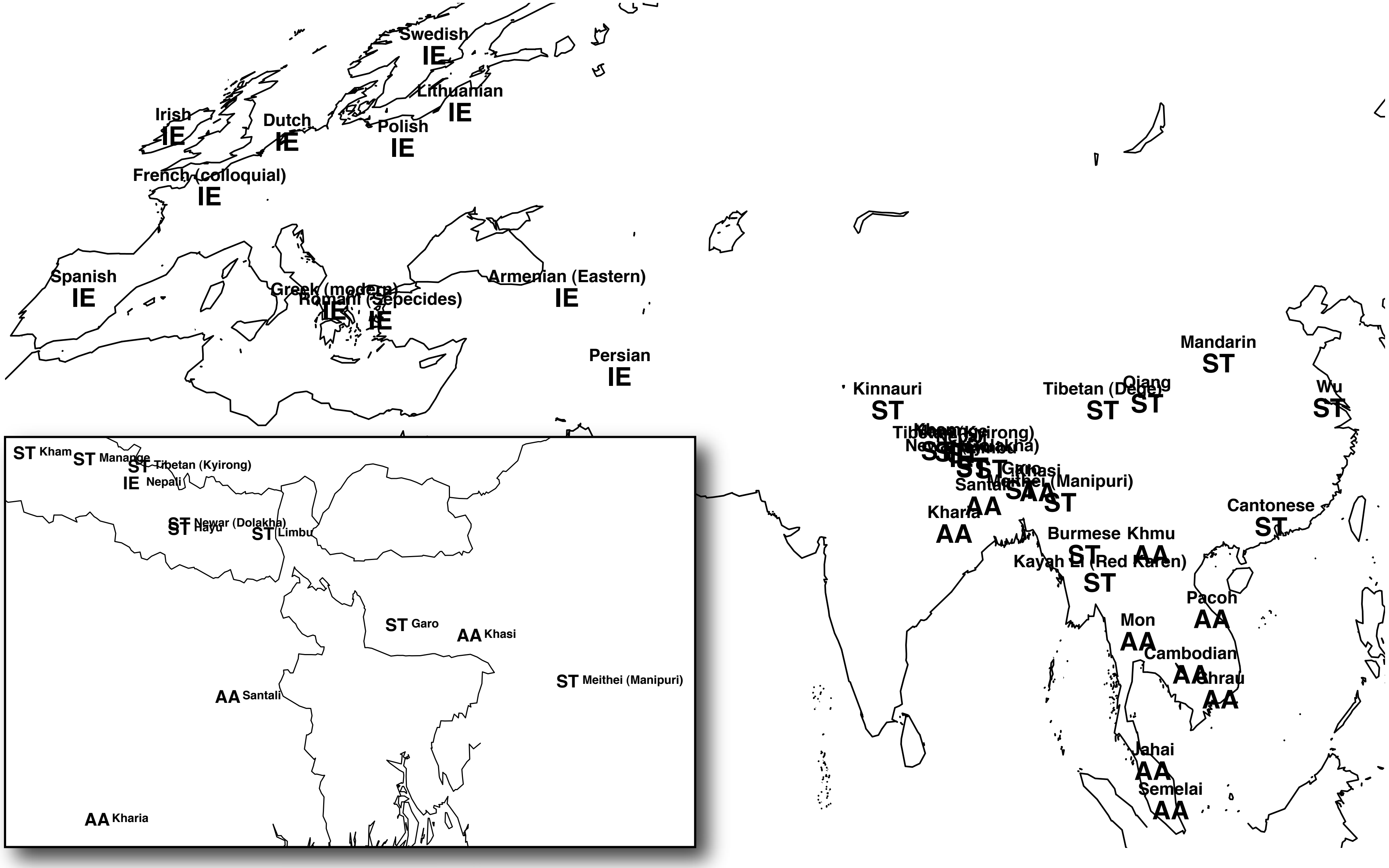
Southeast Asia (Matisoff 2001, Enfield 2005), South-Southwest Asia (Masica 1976, 2001, Ebert 2001); Europe (Dahl 1990, Haspelmath 1998, 2001, Heine & Kuteva 2006)



Bickel, Hildebrandt, & Schiering 2009

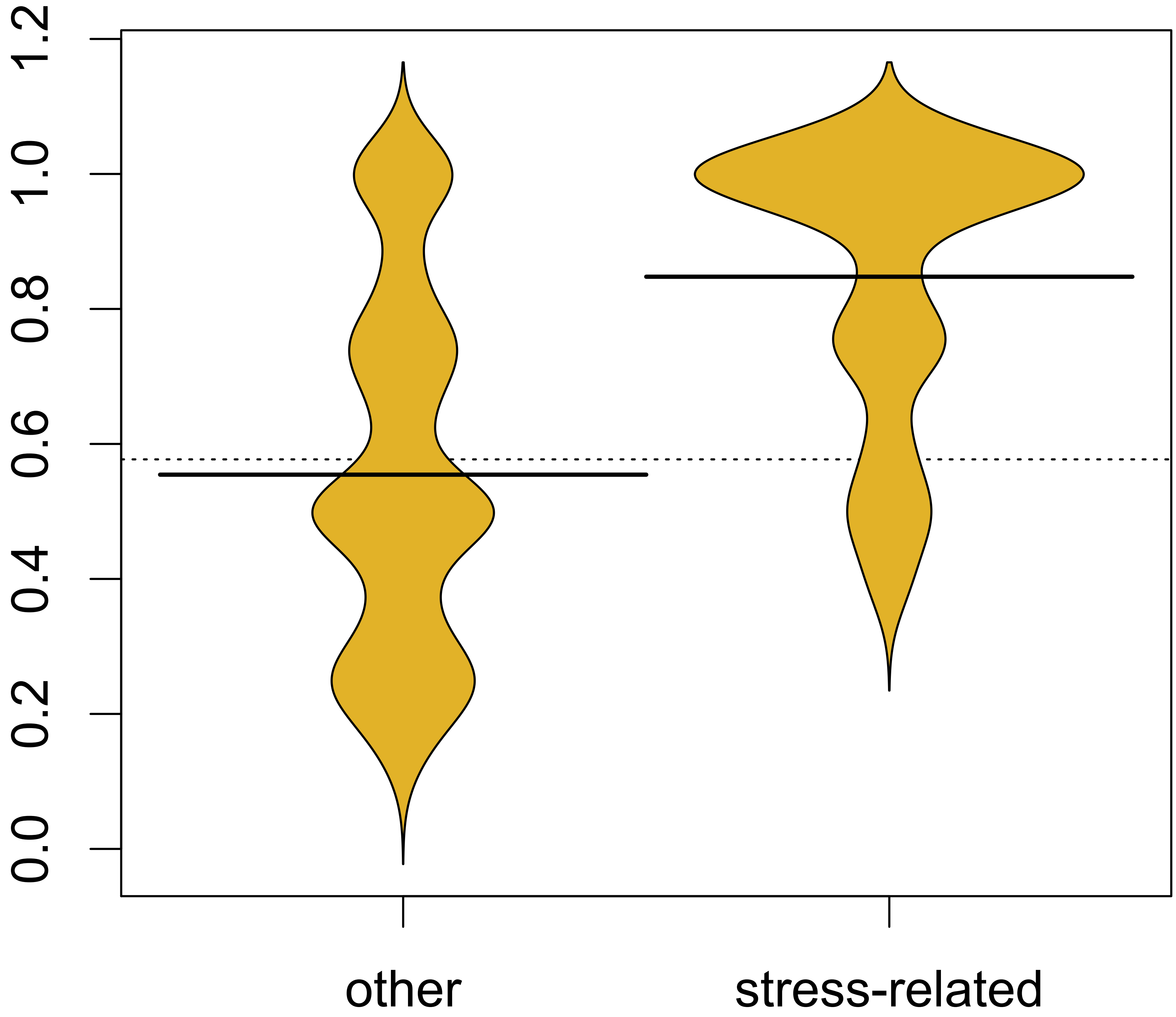
An example: relative p-word size

The chosen areas allow similarly-sized family samples, with one representative per sub-branch of major branches in three families (or two if phonologies are known to be diverse and data are sufficient): Austroasiatic (11), Indo-European (12), Sino-Tibetan (17)



Bickel, Hildebrandt, & Schiering 2009

An example: relative p-word size



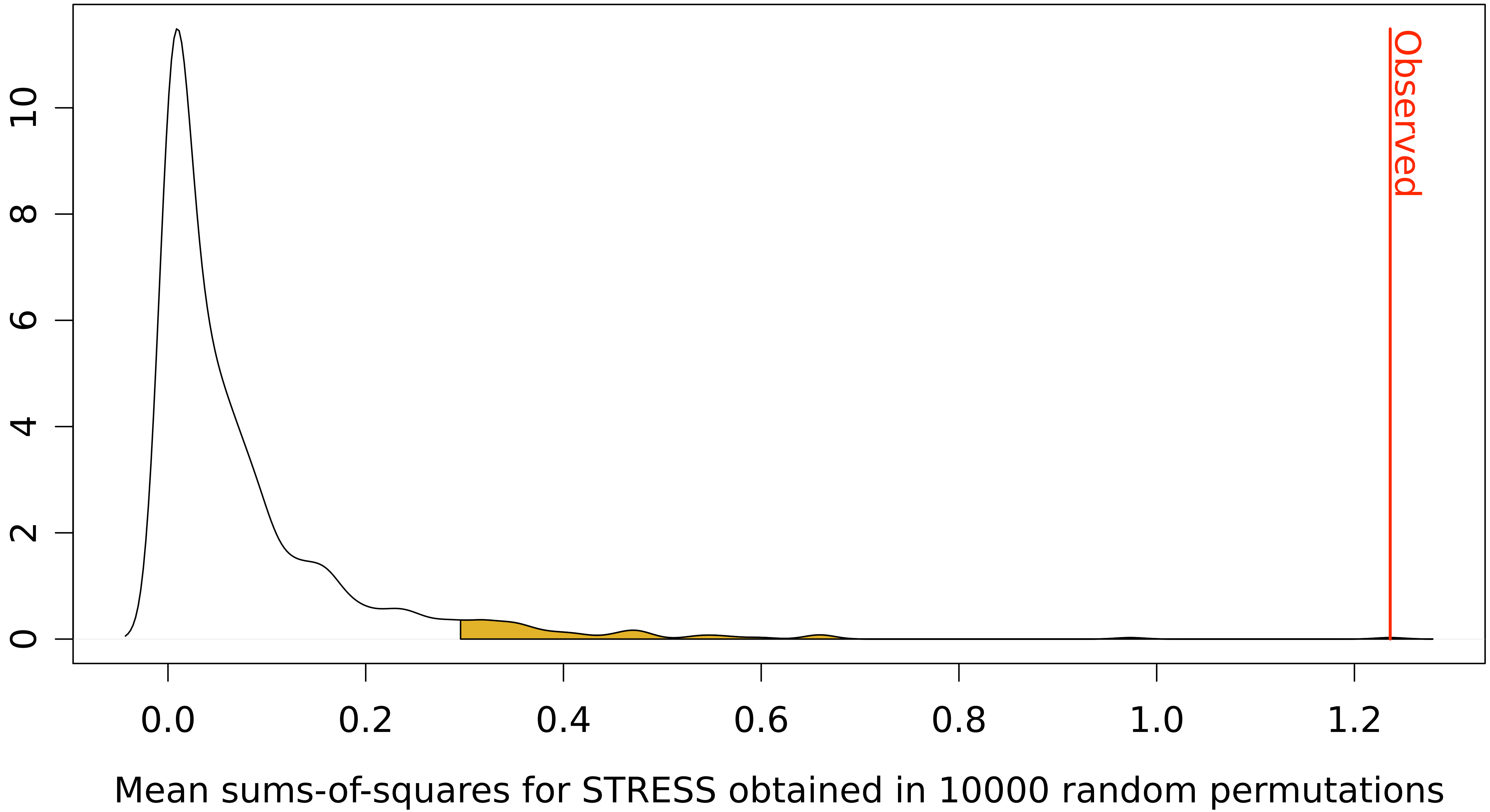
Model testing

The sample is not random; therefore we test the model by Conditional (aka (Approximatively) Exact) Inference — specifically, Monte-Carlo permutation of the response — rather than via Random Sampling Theory (Janssen, Bickel & Zúñiga 2006), e.g.

c	STRESS	STOCK	AREA	Permutation 1	Permutation 2
0.50	other	Indo-European	Indic	0.67	0.20
0.25	other	Indo-European	Indic	0.50	1.00
1.00	stress-related	Sino-Tibetan	Indic	0.50	1.00
0.60	other	Indo-European	Europe	0.25	0.29
0.20	other	Indo-European	Europe	0.86	0.50
0.40	other	Indo-European	Europe	1.00	1.00
0.20	other	Indo-European	Europe	0.25	0.67
0.50	other	Indo-European	Indic	0.20	0.50
0.50	other	Indo-European	Indic	0.25	1.00
1.00	other	Sino-Tibetan	Indic	0.50	1.00
0.75	other	Sino-Tibetan	Indic	0.67	0.50
MS	1.24			0.50	0.03

Bickel, Hildebrandt, & Schiering 2009

Model testing



Model testing

Findings based on how often MS of a given term show up as equal or higher in 10,000 permutations, starting with

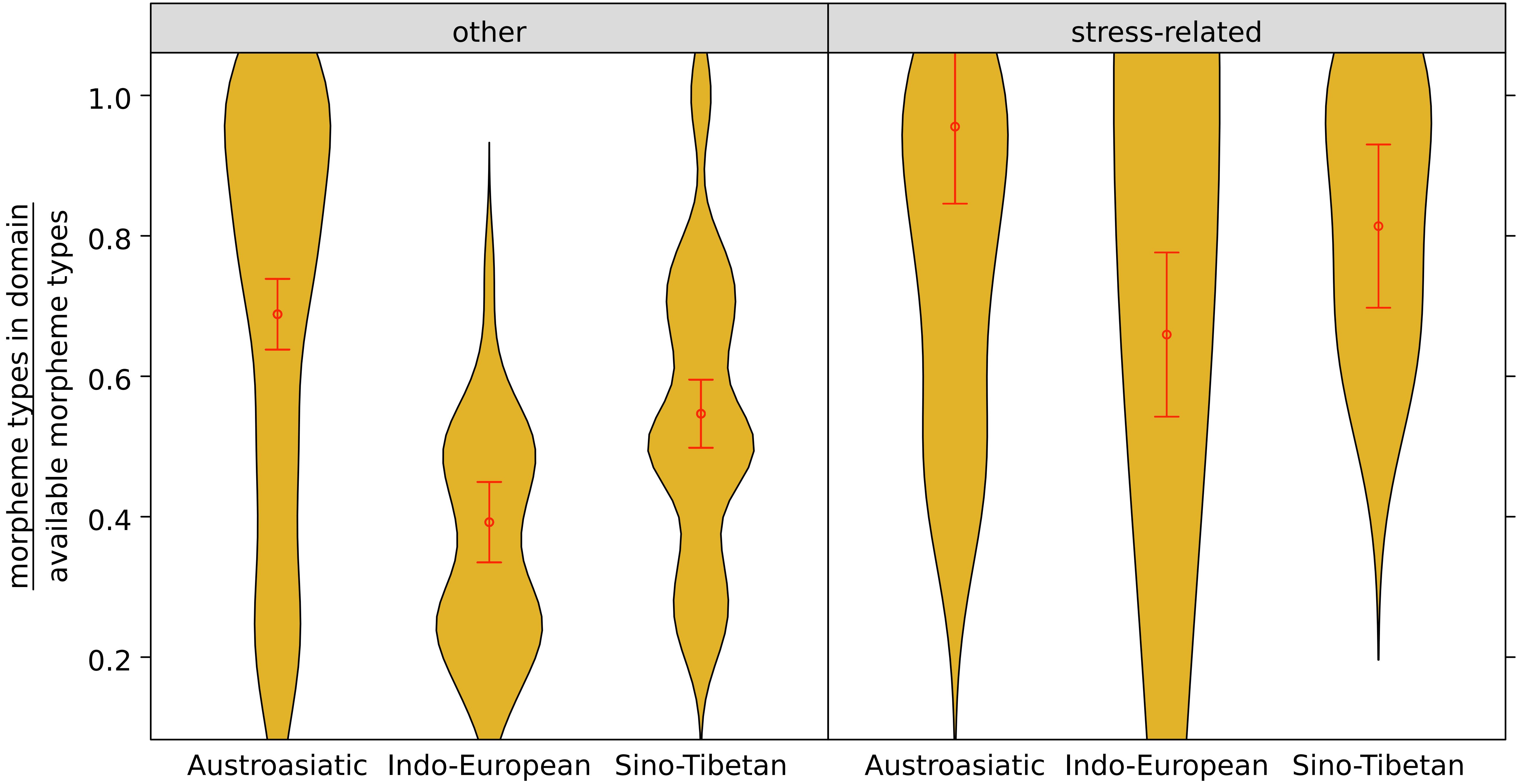
$$\mu(c) = \alpha + \beta_1[\text{STRESS}] + \boldsymbol{\beta_2}[\text{FAMILIES}] + \boldsymbol{\beta_3}[\text{AREAS}] + \boldsymbol{\beta_4}[\text{FAMILIES}][\text{STRESS}] + \boldsymbol{\beta_5}[\text{AREAS}][\text{STRESS}]$$

- [AREAS][STRESS]: $\hat{\beta}_{5.1}=.08$, $\hat{\beta}_{5.2}=.14$, $MS=.008$, ns.
- [FAMILIES][STRESS]: $\hat{\beta}_{4.1}=-.46$, $\hat{\beta}_{4.2}=-.19$, $MS=.13$, ns.
- [AREAS]: $\hat{\beta}_{3.1}=.1$, $\hat{\beta}_{3.2}=.06$, $MS=.08$, n.s.
- [FAMILIES]: $\hat{\beta}_{2.1}=-.30$, $\hat{\beta}_{2.2}=-.14$, $MS=1.67$, $p<.001$
- [STRESS]: $\hat{\beta}_1=.26$, $MS=1.18$, $p<.001$
- Therefore, the most parsimoneous model is...

Bickel, Hildebrandt, & Schiering 2009; Janssen, Bickel, & Zúñiga 2006

Model testing

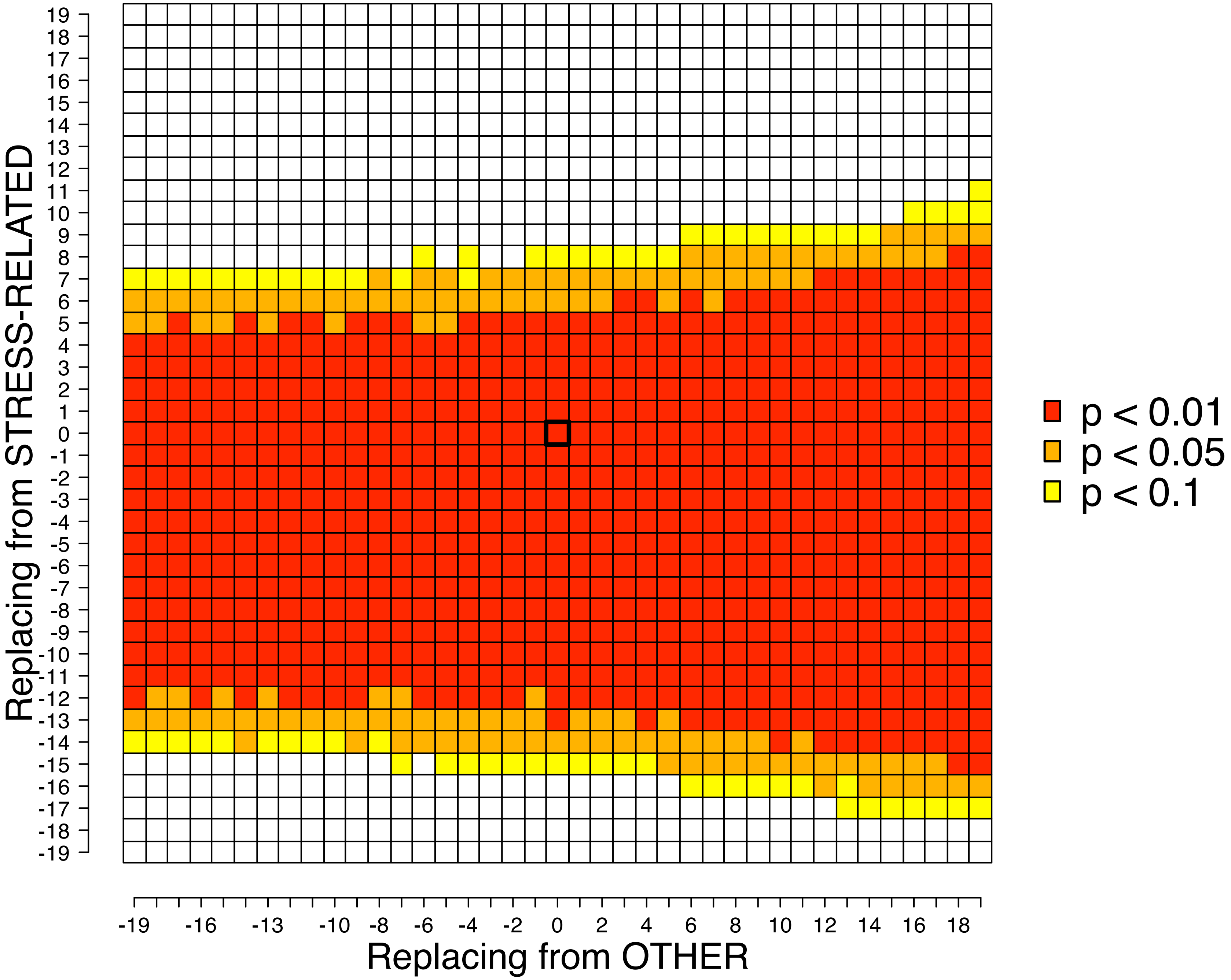
$\hat{\mu}(c) = .69 - .30IE - .14ST + .26STRESS$



Note: factors were parametrized as follows: IE: ±Indo-European (the diff. between non-IE and IE), ST: ±Sino-Tibetan (the diff. betw. non-ST and ST), so -IE - ST (i.e. IE=0 & ST=0) means Austroasiatic (i.e. the baseline); STRESS is the diff. between non-stress and stress

Reliability estimation

Since there are many less stress-related pw-patterns (19) than others (222), we also need a Reliability Analysis (Janssen et al 2006), replacing critical values of *c* by their grand mean:

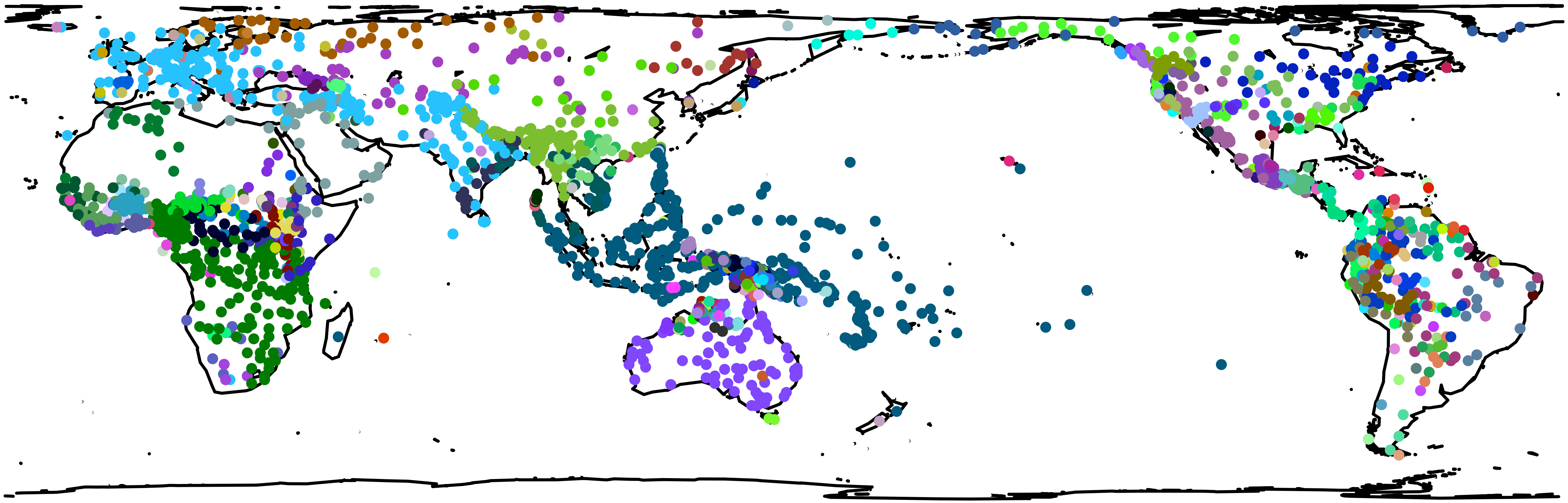


Interim conclusions

- The universal is independent of the effects from language contact (areas) and blind inheritance (families), but
  - blind inheritance **also (independently)** matters — but without interaction!
  - This is in line with known diachronic preferences for structure preservation (Blevins 2004)

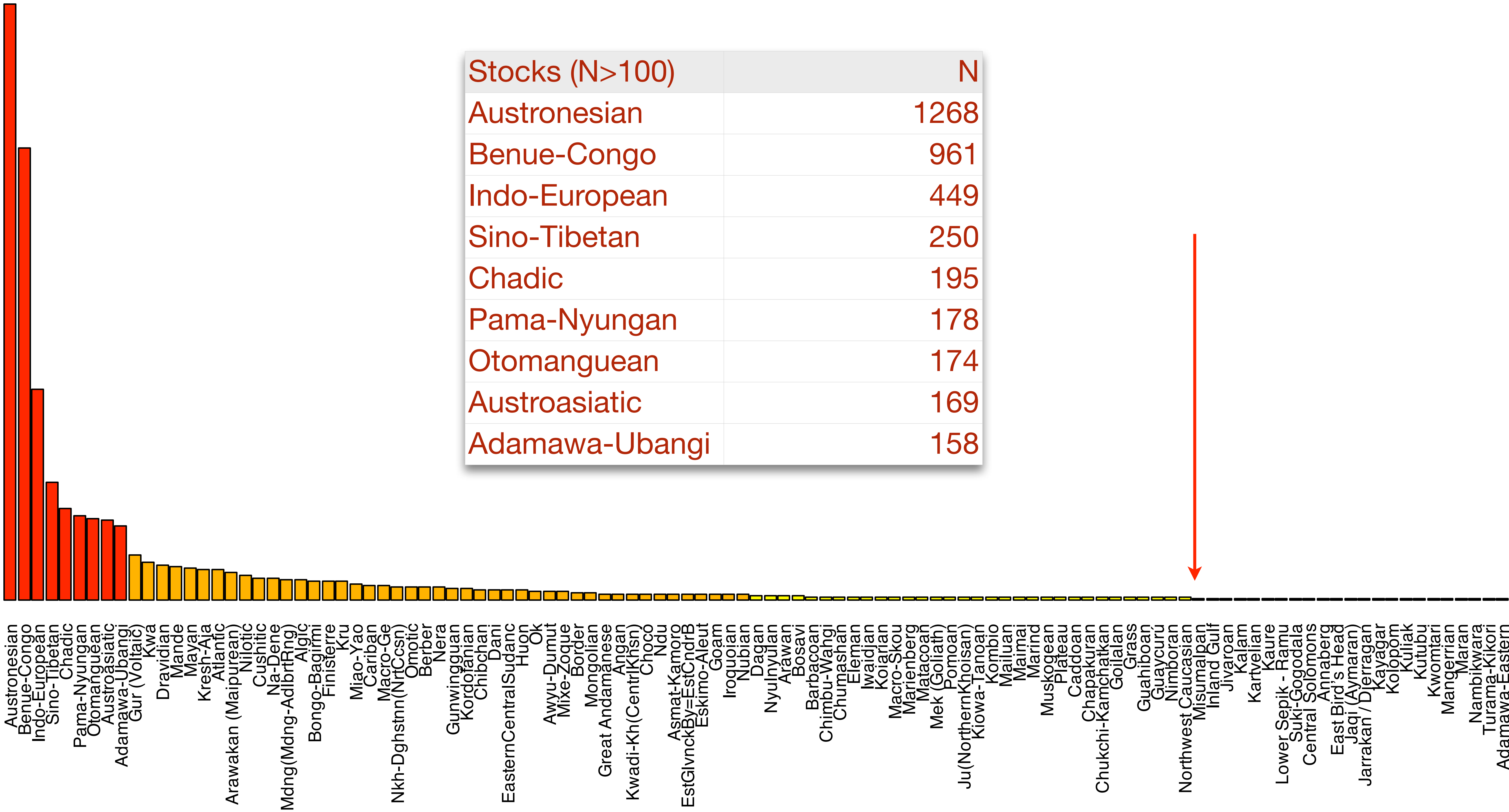
Prospects for larger datasets

- But… how can we extend the test to a worldwide database?
- Including the stock factor into regression models is problematic in worldwide datasets because there are over 300 stocks, ca. 40% having only one member



Prospects for larger datasets

Stocks (N>100)	N
Austronesian	1268
Benue-Congo	961
Indo-European	449
Sino-Tibetan	250
Chadic	195
Pama-Nyungan	178
Otomanguean	174
Austroasiatic	169
Adamawa-Ubangi	158



Towards a new approach

- Three observations:
  1. Linguistic structures change over time.
  2. Sets of changes over time is what defines reconstructable families.
  3. Within these sets, each change can be affected by a factor of Structural Pressure or Language Contact (either by militating against or pushing for change)
- If many such sets are affected in the same way, this suggests that the factor leads to a specific bias in many families.
- The stronger the factor, the more families will be bias and relatively homogenous

Towards a new approach

- Therefore, all universals are in fact diachronic in nature (Greenberg 1978, 1995, Maslova 2000, Nichols 2003, etc.)!
    - We can reformulate every universal as diachronic pressure, e.g. “VO → NRel” can be reformulated as: $\pi(\text{VO\&RelN} \succ \text{VO\&NRel}) > \pi(\text{VO\&NRel} \succ \text{VO\&RelN})$, where “$\succ$” symbolizes diachronic change
    - Given this, ‘diachronic universals’ are just a special case, concerning the explanation of the universal

      “OV → Np”: $\pi(\text{OV\&pN} \succ \text{OV\&Np}) > \pi(\text{OV\&Np} \succ \text{OV\&pN})$ because N_OV frequently develops into Np (e.g. Nepali *āja bhane* ‘today (N_O) saying (V)’ > ‘as for (P) today (N)’)

Towards a new approach: the Family Bias Method

- We can estimate diachronic pressures on (non-singleton) families by assessing whether or not they show a specific bias:
  - Expectation if there was no pressure at work:
    - if a variable is stable (‘blind inheritance’), families end up biased in whatever way the proto-language happened to be bias, ***under whatever other conditions***
    - If a variable is instable (random fluctuation, unknown factors), families tend to diversify over time
  - Expectation if there was universal pressure at work:
    - families tend to show the same bias (the one proposed by the universal), whatever their proto-language, and regardless of area and random fluctuation

Bickel 2008, 2010 [Mss.], book in preparation

Towards a new approach: the Family Bias Method

- Therefore, if most (according to some statistical test) families are biased in the same way regardless of their areal locations and regardless of any structural condition, this attests to universal pressure.
- To what extent is this a valid inference?
- Two models to be contrasted:

  A. Family biases reflect universal trends:

  $\pi(\text{non-F} \succ \text{F}) > \pi(\text{F} \succ \text{non-F})$

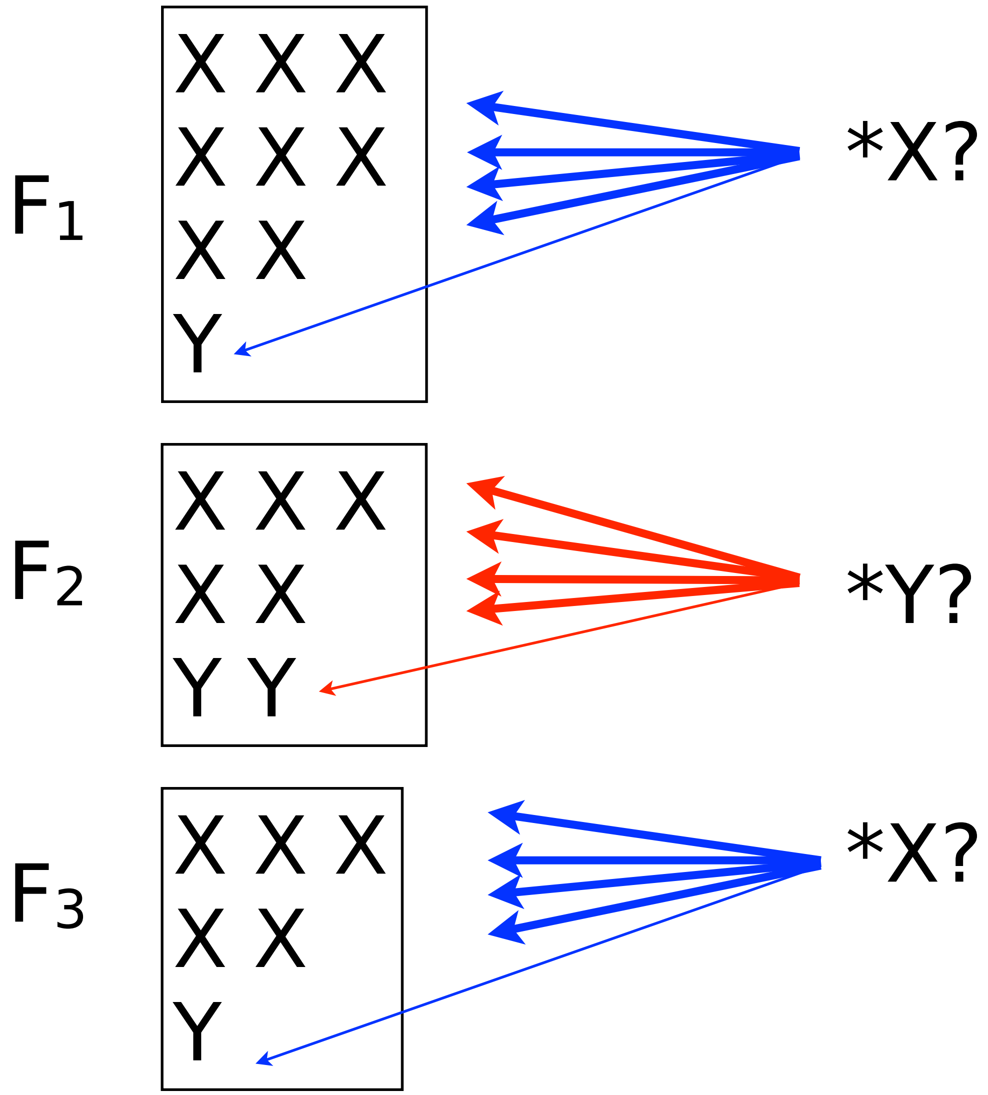
  B. Family biases reflect extreme stability:

  $\pi(\text{non-F} \succ \text{F}) \approx \pi(\text{F} \succ \text{non-F}) \approx 0$ (Maslova 2000)

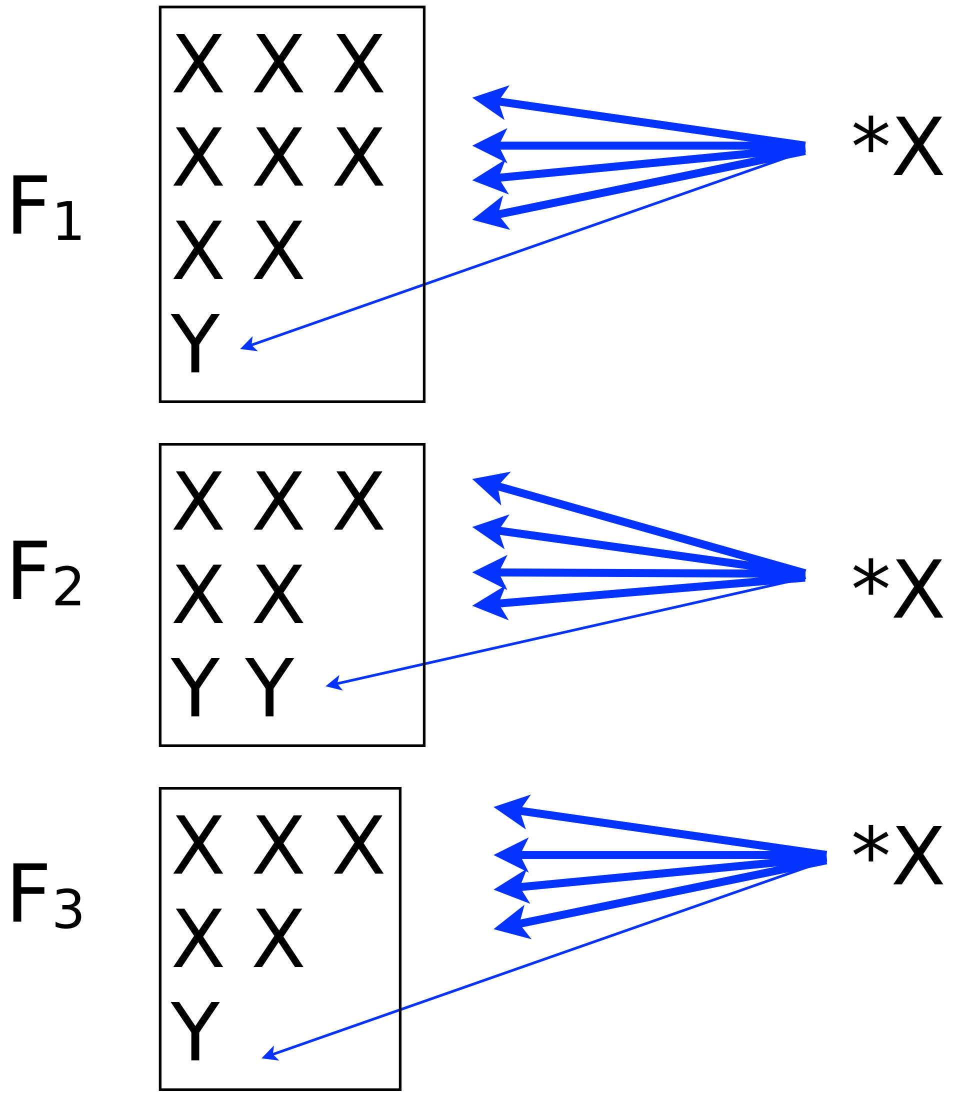
Towards a new approach: the Family Bias Method

Illustration: most families are biased in the same way, towards “X”

Interpretation A:
$\pi(Y \succ X) > \pi(X \succ Y)$



Interpretation B:
$\pi(Y \succ X) \approx \pi(X \succ Y) \approx 0$



Can we tell the difference?

Is $\pi \approx 0$ (extreme stability) plausible?

- Assume $\pi \approx 0$ for a variable distributed as *D*.
- Then, $D(G_0)$ must reflect $D(G_{i+1})$.
- Unless there was universal pressure before G_{i+1}, all $D(G_i)$ must reflect $D(G_{i+1})$ until *i* spans the entire history of the human language faculty.
- Then, changes in *D* are extremely unlikely within short time intervals.
- Assume that all reconstructible time intervals are relatively short (up to about 8Ky, the age of provable families)
- Expect to be able to observe almost no changes in $D(G_0)$.
- Is this so? Given a set of variables, how many of them show changes in $D(G_0)$?

Bickel forthcoming (“Distributional biases in language families”) 33

Is $\pi \approx 0$ (extreme stability) plausible?

- The minimum number of attested changes C for a variable V with k *attested* types (“levels”) in a family F is

  $\min(C_F) = k_F - 1$

A family with 10 languages: A A A A A B B B A A, so $k_F = 2$
Minimum change scenarios:
  *A ≻ B in one branch, the rest stays, *or*
  *B ≻ A in one branch, the rest stays

Another family with 10 lgs: A A C A A B B B A A, so $k_F = 3$
Minimum change scenarios:
  *A → B in F_1, *A → C in F_2, A stays in F_3 or
  *B → A in F_1, *B → C in F_2, B stays in F_3, or
  *C → A in F_1, *C → B in F_2, C stays in F_3

That’s the logical minima. (There can always many more!)

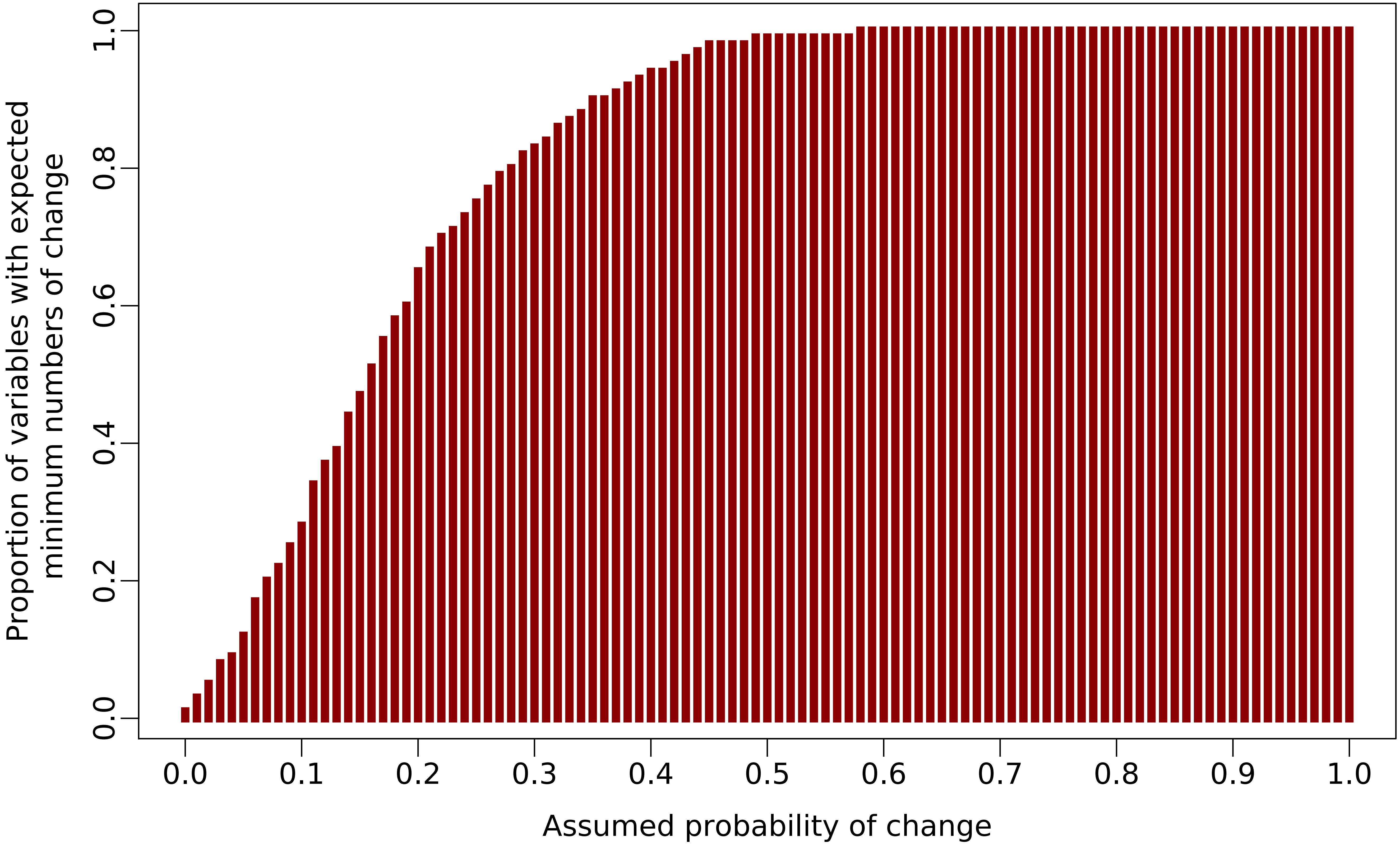
Is $\pi \approx 0$ (extreme stability) plausible?

- 448 per-language variables from WALS and AUTOTYP, with data for at least 10 non-singleton families
  - all kinds of structure
  - many recodings (e.g. both a 6-way word order typology and a binary OV vs VO coding)
  - all variables treated as categorical because all changes are categorical
- Test for each variable whether the observed minimum of changes per family **exceeds what can be expected** under some assumed probability of change π, *and no other factors.*

Is $\pi \approx 0$ (extreme stability) plausible?

- **Criterion of excess:** the proportion of min(C_F) out of the total minimum of opportunities O_F for change is unexpected for a given π if the prop. ≥ the prop. under H_0 in a binomial test (at $\alpha = .05$)
  - opportunities for change = $(k\text{-}1)\cdot N$(families), e.g.
    - assume k=2 (*logically possible types*) and 50 families,
      - then there are *at least* 50 opportunities for change
      - finding 20 out of 50 is unexpected under $\pi = .15$ (at a α =.05 rejection level) → “excess”
    - assume k=3 (*logically possible types*) and 50 families,
      - then there are at least 2 opportunities of change per family (e.g. A ≻ B and B ≻ C), i.e. total of at least 100 opportunities
      - finding 20 out of 100 is expected under $\pi = .15$

Is $\pi \approx 0$ (extreme stability) plausible?



448 variables

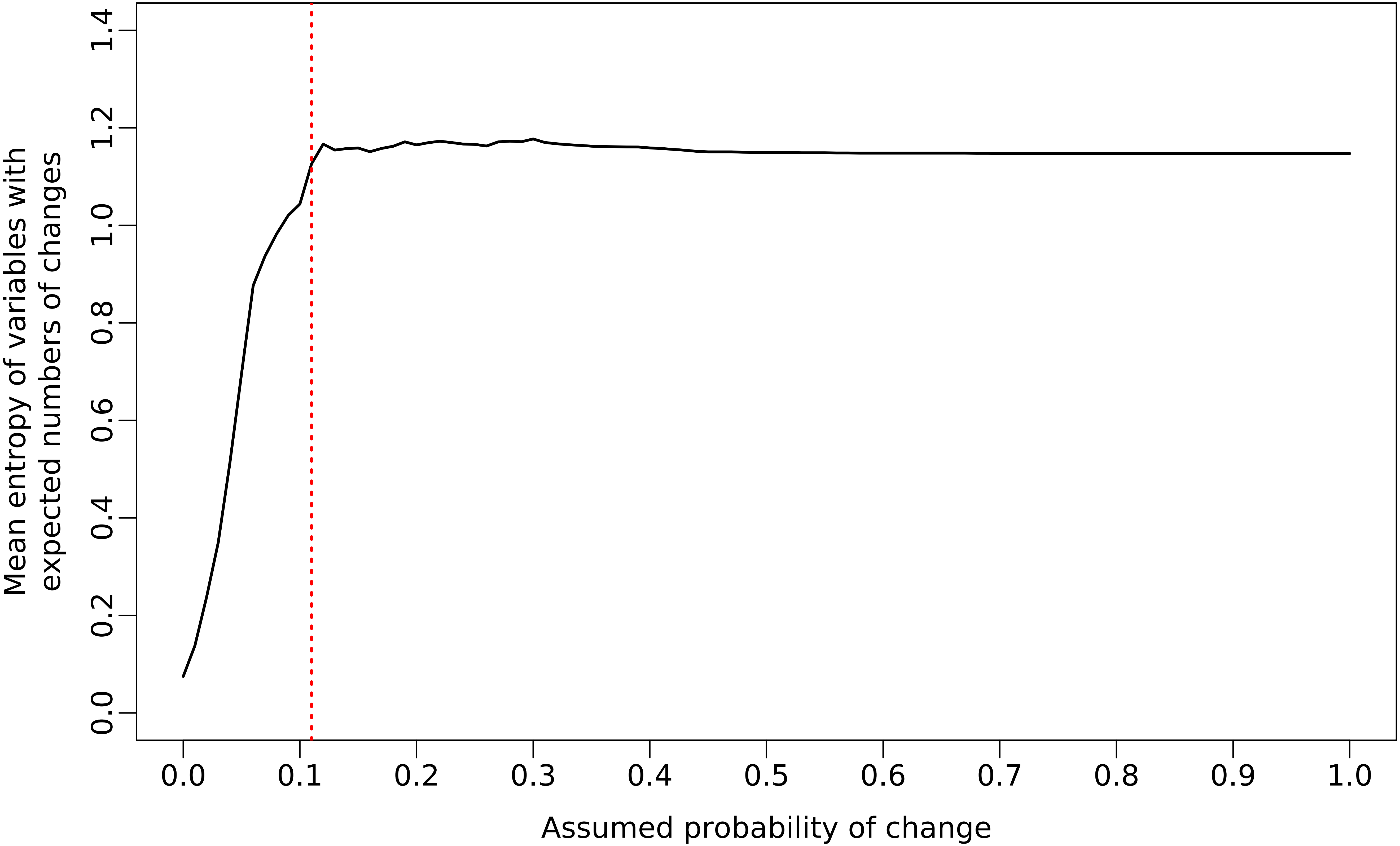
Is $\pi \approx 0$ (extreme stability) plausible?

- A closer look at the variables for which we find no more changes than a small π lets us expect
- Estimate the information entropy of these variables, as a measure of bias.
- Example: some of the variables with min(C_F) under $\pi = .01$:

Variable (and data source)	Changes N_{min}	Opportunities N_{min}	Entropy $\hat{H}$	Ratio of values
Interrog./decl. distinction (Dryer, 2005a)	1	89	0.01	841:1
Indep. subject pronouns (Daniel, 2005)	0	31	0.07	258:2
Tonal case (autotyp and Dryer, 2005b)	3	91	0.07	698:6
Stem flexivity condit. by NEG (autotyp)	0	40	0.12	141:1:1
'Have'-perfect (Dahl & Velupillai, 2005)	1	15	0.35	101:7
Co-exponent type of NEG (autotyp)	4	234	0.60	185:5:3:1:1:1:1:1:1:1

- This is typical: $\pi \leq .10$ suggest **rara vs. universalia** distributions, not extreme stability

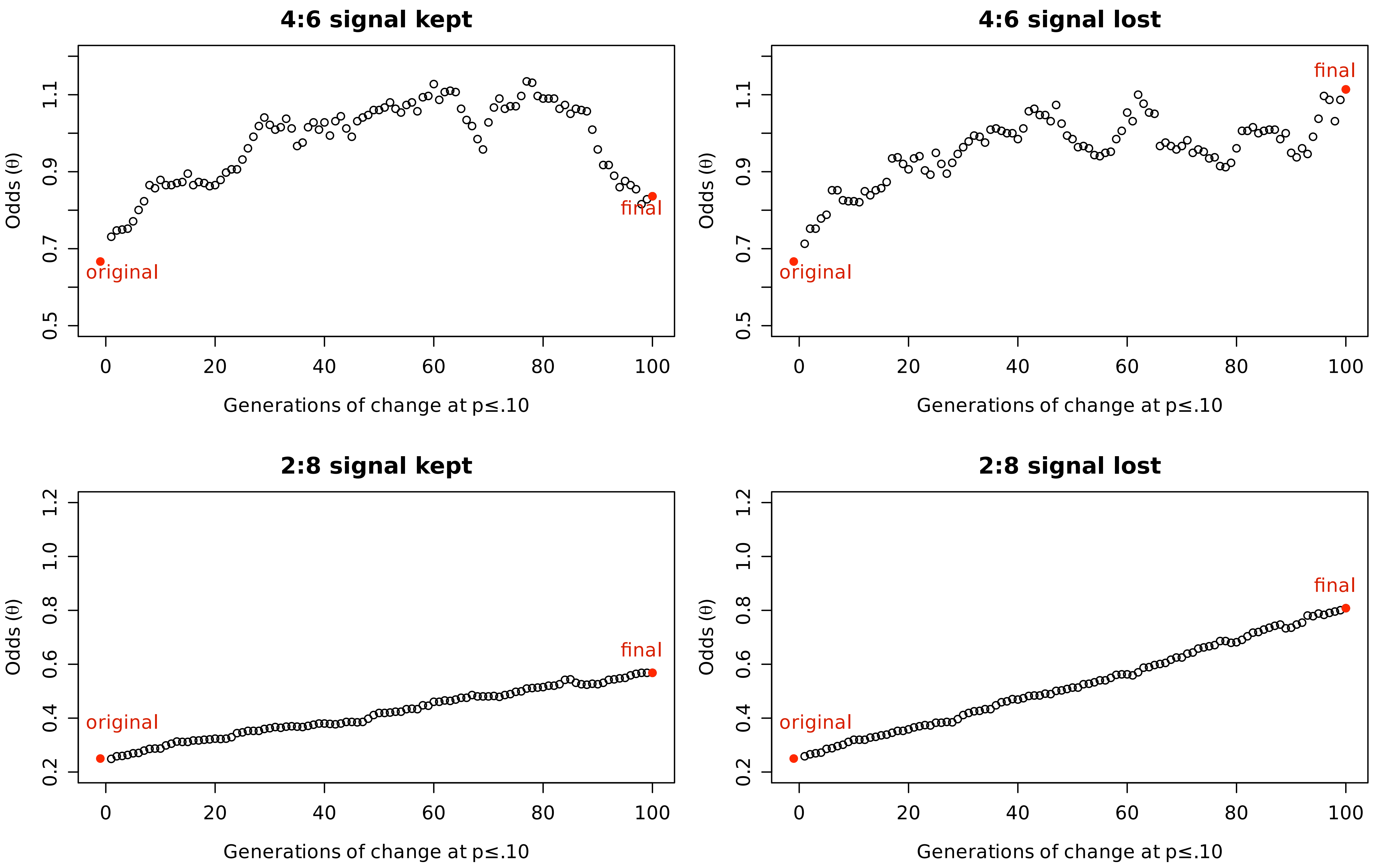
Is $\pi \approx 0$ (extreme stability) plausible?



Further evidence against small π

- Already with $\pi = .05$ typological distributions change so much that they can no longer be detected after 100 generations of change, which is a reasonable minimum age of the human language faculty:
- Simulations: take some initial distribution of 1,300 languages (e.g. with 80% A and 20% B) and let randomly change between 0 and π languages over 100 generations
  - 1,300 corresponds to the largest datasets we know (Dryer 2005)
  - random changes between 0 and π because π is not constant. (This favors evidence for stability!)
- Examples:

Further evidence against small π: sample simulations



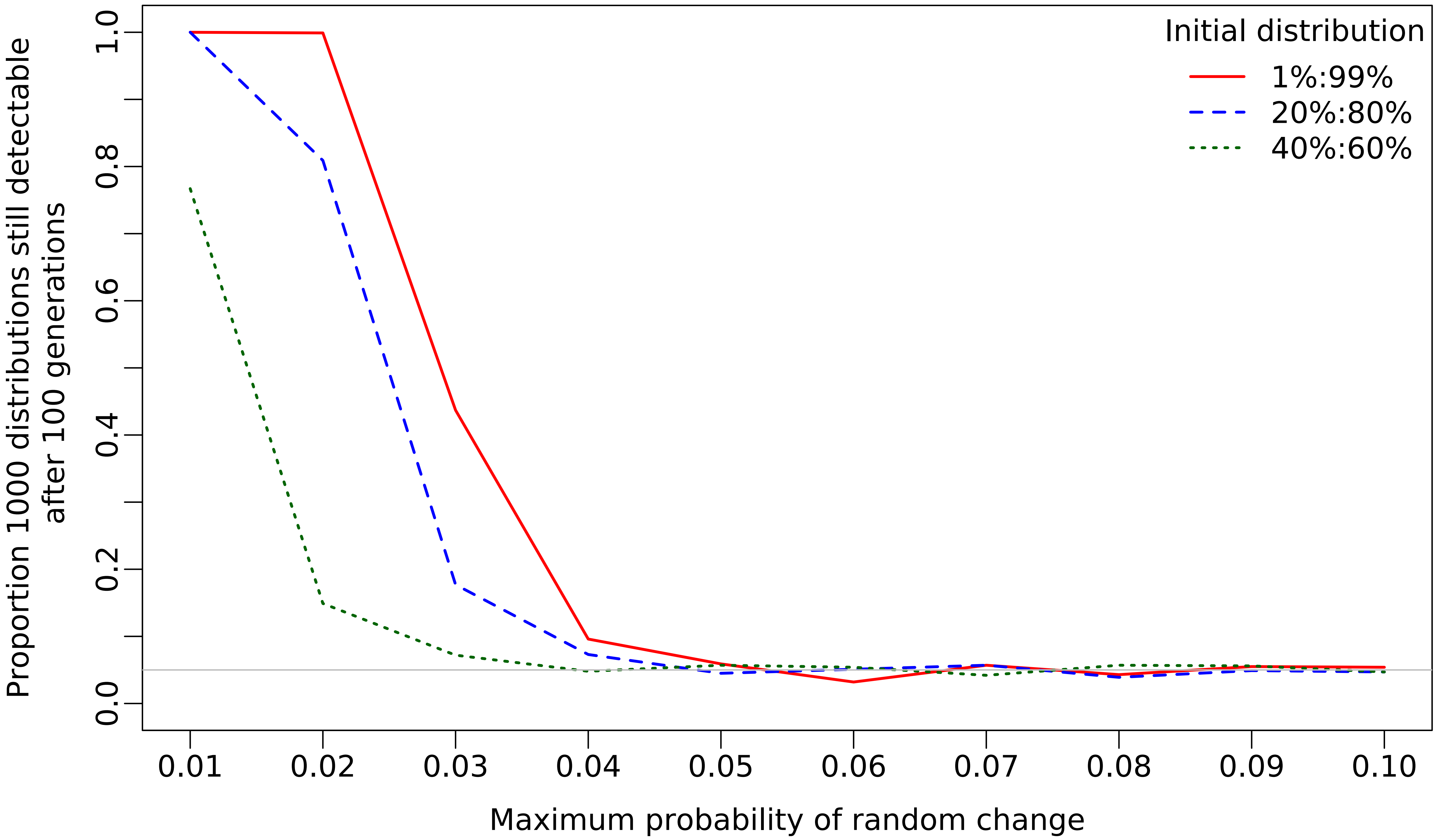
4:6 signal kept

4:6 signal lost

2:8 signal kept

2:8 signal lost

Further evidence against small π: results



Interim conclusions

- Assuming small π is not compatible with the number of changes that we observe in most variables
  - small π matches what we observe only for less than 10% of known variables ($N = 448$)
  - variables that match small π suggest rara vs. universalia distributions (such as tonal case), not stability
- Even $\pi = .05$ is too high for distributions to persist over 100 generations.

Family biases as signals of universal and areal trends

- These findings suggest that if we find **consistent distributional biases across families**, these result from

  $\pi(\text{non-F} \succ \text{F}) > \pi(\text{F} \succ \text{non-F})$, i.e. **pressure over time**

  and **not from extreme stability** (*pace* Maslova 2000),

  and, therefore, **cannot be taken as estimates of diachronic stability/inertia** (*pace* Parkvall 2008, Wichmann & Holman 2009)

Family biases as signals of universal and areal trends

A closer look at what has been hypothesized — erroneously, I think! — to be very stable on the basis of family biases:

Parkvall’s Top Ten:

1. Phonemic clicks
2. Verbal property predicates
3. M-T pronouns
4. Suffixal plural in nouns
5. [N N-gen] order
6. medium synthesis degree
7. ‘black’ = ‘blue’
8. **[N p] order**
9. [p N] order
10. **[N-gen N] order**

Wichmann & Holman’s Top Ten

1. Syncretism in subject agreement
2. Postposed dummy heads (‘a red one’)
3. Relative Pronouns (for Subj. and Obliques)
4. 8 or between 8 and 9 color categories
5. Non-sex-based gender system
6. No inflectional case marking
7. Nonverbal property predicates
8. Sex-based gender system
9. No inclusive/exclusive opposition
10. **[N p] order**

Parkvall 2008, Wichmann & Holman 2009

Family biases as signals of universal and areal trends

- Interestingly, Wichmann & Holman 2008 note that the more stable a feature is, the more it correlates with others (accounting for 7-19% of the variance, depending on how one controls for genealogical relatedness)
- yes, because ‘stable’ here means ‘having a strong and uniform family bias’ and
- ‘having a strong and uniform family bias’ is evidence for a universal, here a correlation between features, e.g.

Np ~ [N-gen N] oder!

Family biases as signals of universal and areal trends

Thus, trends in family biases suggest pressure over time, as a result of

a. universal principles if the trend is the same in all areas under the same conditions

b. areal diffusion if the trend varies significantly from area to area

Case study: S=A preference

- ERP studies of several languages (English, German, Italian, Turkish, Japanese, Hindi) suggest that the human processor prefers
  - initial NPs with a default case (marked or zero, but assigned by the majority class of verbs) to refer to S or, if that doesn’t work, A.

    (presumably because these interpretations require less complex structures: Bornkessel-Schlesewksy & Schlesewksy 2006, 2009, Wang et al. 2009, 2010, Choudhary et al. 2010, Wolff et al. 2008)
  - This constantly invites a reanalysis of default cases as S=A cases, disfavoring ergativity (S≠A)

Case study: S=A preference

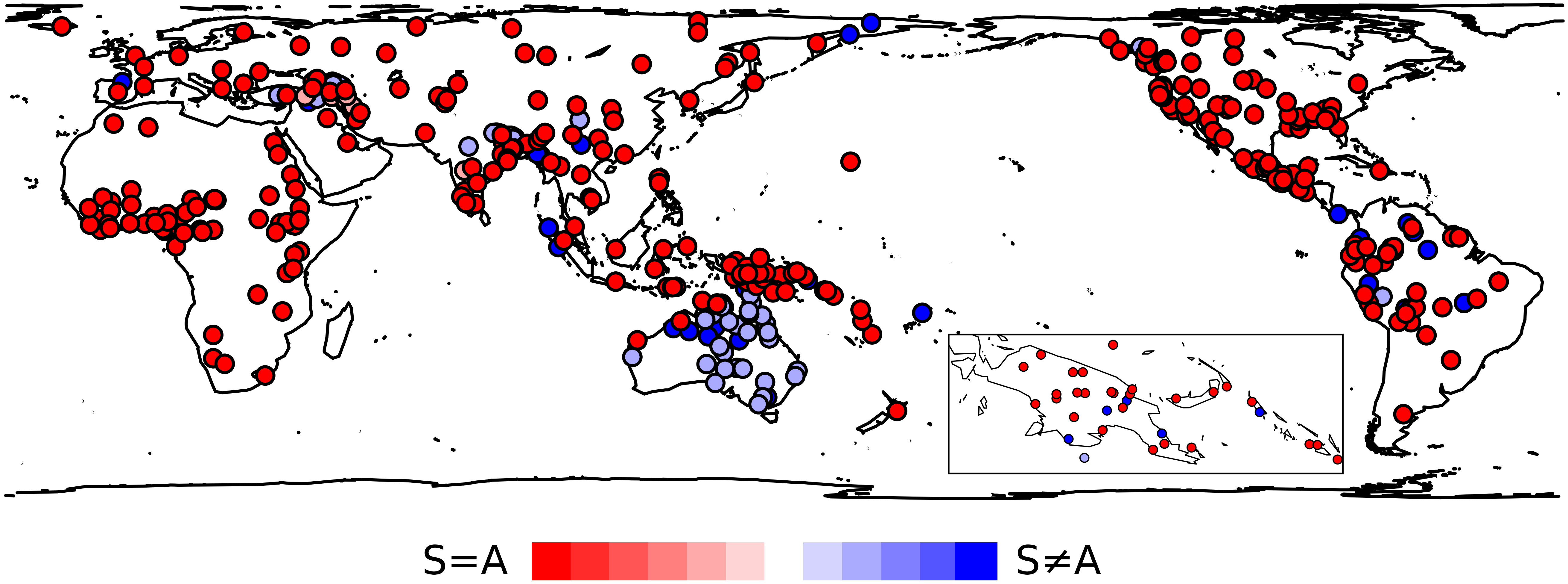
- *Hypothesis:* there is a universal bias for S=A and against S≠A (i.e. a universal Bias Against Ergativity),
  - at least for non-V-initial languages
  - and for cases assigned by the default predicates
- Test this using the Family Bias Method

Case study: S=A preference

- AUTOTYP database developed by Alena Witzlack-Makarevich and myself, coding the sets of roles covered by each case marker (“case frames”) under any conditions that a language may have (multivariate!)
    - N = 411 languages
- select the case frames assigned by default predicate classes (as per the hypothesis), separately per condition (e.g. past tense vs. non-past tense, first/second vs. third person etc.)
    - N = 489 individual frames
- compute whether the frames align S=A vs S≠A

Case study: S=A preference

Proportion of S=A case alignments per language (across all known splits and conditions):



- Need to control for confounding effects especially from the Australia/New Guinea and Eurasia macroareas
- 4-way-breakdown of world as a control factor

Applying the Family Bias Method: Step 1

a. Find the highest genealogical taxa

$t \in$ {stock, major branch, language, individual case system}

that are not split across the 4 macroareas and compute the proportions of S=A alignments within t

b. Estimate the diachronic bias of t by binomial tests on whether the proportion of S=A exceeds what can be expected under random data permutation (with a relatively high rejection level $\alpha = .10$ because of discreteness and power loss problems in small stocks). Possible outcomes:

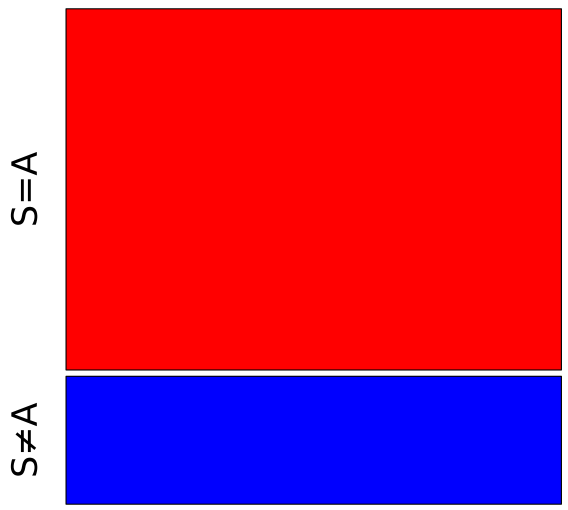
- Bias towards S=A
- Bias towards S≠A
- No bias (diverse)

- **NOTE: it doesn't matter at what taxonomic level we find biases! There is no assumption about time depth!**

Step 1: distribution within families

- Example: Indo-European

mbranch	language	ACC
Albanian	Albanian	S=A
Anatolian	Hittite	S=A
Anatolian	Hittite	S≠A
Balto-Slavic	Russian	S=A
Balto-Slavic	Bulgarian	S=A
Germanic	English	S=A
Germanic	German	S=A
Germanic	Icelandic	S=A
Greek-Armenian	Armenian (Eastern)	S=A
Indo-Iranian	Hindi	S≠A
Indo-Iranian	Hindi	S=A
Indo-Iranian	Maithili	S=A
Indo-Iranian	Persian	S=A
Indo-Iranian	Assamese	S=A
Indo-Iranian	Kurmanjî	S≠A
Indo-Iranian	Kurmanjî	S=A
Indo-Iranian	Marathi	S=A
Indo-Iranian	Marathi	S≠A
Indo-Iranian	Marathi	S=A
Indo-Iranian	Tajik	S=A
Indo-Iranian	Rushan	S=A
Indo-Iranian	Rushan	S=A
Indo-Iranian	Rushan	S≠A
Indo-Iranian	Ishkashimi	S=A
Indo-Iranian	Vafsi	S≠A
Indo-Iranian	Vafsi	S=A
Indo-Iranian	Vafsi	S=A
Indo-Iranian	Yazgulyam	S=A
Indo-Iranian	Yazgulyam	S≠A
Indo-Iranian	Yazgulyam	S=A
Indo-Iranian	Khufi	S=A
Indo-Iranian	Khufi	S=A
Indo-Iranian	Khufi	S≠A
Indo-Iranian	Shughni	S=A
Indo-Iranian	Bartangi	S=A
Indo-Iranian	Bartangi	S≠A
Indo-Iranian	Bartangi	S=A
Indo-Iranian	Avestan	S=A
Indo-Iranian	Kirmanjki	S=A
Indo-Iranian	Kirmanjki	S=A
Indo-Iranian	Kirmanjki	S≠A
Indo-Iranian	Dimili	S=A
Indo-Iranian	Dimili	S=A
Indo-Iranian	Dimili	S≠A
Indo-Iranian	Tarom	S=A
Indo-Iranian	Tarom	S=A
Indo-Iranian	Tarom	S≠A
Indo-Iranian	Kajali	S=A
Indo-Iranian	Kajali	S=A
Indo-Iranian	Kajali	S≠A
Indo-Iranian	Shahrudi	S=A
Indo-Iranian	Shahrudi	S=A
Indo-Iranian	Shahrudi	S≠A
Indo-Iranian	Chali	S≠A
Indo-Iranian	Chali	S=A
Indo-Iranian	Chali	S=A
Indo-Iranian	Eshtehardi	S=A
Indo-Iranian	Eshtehardi	S=A
Indo-Iranian	Eshtehardi	S≠A
Indo-Iranian	Talysh (Northern)	S=A
Indo-Iranian	Talysh (Northern)	S≠A
Indo-Iranian	Talysh (Northern)	S=A
Italic-Celtic	Spanish	S=A
Italic-Celtic	Catalan (Standard)	S=A
Italic-Celtic	French, Cajun	S=A



74% S=A
One-sided Binomial test against a 50:50 distribution
p < .0001

Step 1: distribution within families

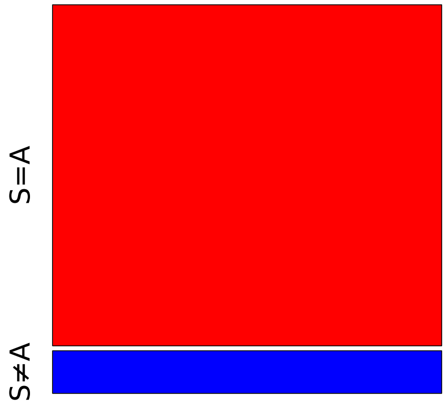
- Sometimes area distinctions (or distinction made by other control factors) cross-cut families.
  - In many cases we can just move down a bit in the genealogy so that subgroups are fully contained within the relevant areas.
  - Moving down the genealogy is OK because we are interested in detecting any bias, at whatever time-depth!

    (But there is a problem in that we get many small families. We will come back to this.)

Example: Austronesian

macrocontinent	mbranch	sbranch	ssbranch	language	ACC
Eurasia	Malayo-Polynesian	**Borneo**		Mualang	S=A
Eurasia	Malayo-Polynesian		**Aceh-Chamic**	Acehnese	S≠A
Eurasia	Malayo-Polynesian		**Aceh-Chamic**	Cham (Western)	S=A
Eurasia	Malayo-Polynesian			**Batak (Toba)**	S=A
Eurasia	Malayo-Polynesian			**Balangao**	S=A
Eurasia	Malayo-Polynesian			**Bontok**	S=A
Eurasia	Malayo-Polynesian			**Ilocano**	S=A
Eurasia	Malayo-Polynesian			**Javanese**	S=A
Eurasia	Malayo-Polynesian			**Nias**	S≠A
Eurasia	Malayo-Polynesian			**Nias**	S≠A
NG-Australia	Malayo-Polynesian	Eastern Malayo-Polynesian		**Seimat**	S=A
NG-Australia	Malayo-Polynesian	**Central Malayo-Polynesian**	Central Maluku	Lamaholot	S=A
NG-Australia	Malayo-Polynesian	**Oceanic**	Central Pacific	Maori	S=A
NG-Australia	Malayo-Polynesian	**Oceanic**	Central Pacific	West Futuna	S=A
NG-Australia	Malayo-Polynesian	**Oceanic**	Southern Oceanic	Dumbea	S=A
NG-Australia	Malayo-Polynesian	**Oceanic**		Arosi	S=A
NG-Australia	Malayo-Polynesian	**Oceanic**		Lonwolwol	S=A
NG-Australia	Malayo-Polynesian	**Oceanic**		Futuna (East)	S≠A
NG-Australia	Malayo-Polynesian	**Oceanic**		Mwotlap	S=A
NG-Australia	Malayo-Polynesian	**Oceanic**		Siar	S=A
NG-Australia	Malayo-Polynesian	**Oceanic**		Iduna	S=A
NG-Australia	Malayo-Polynesian		**Gorontalo-Mongondic**	Gorontalo	S=A
NG-Australia	Malayo-Polynesian		**South Sulawesi**	Bugis	S=A

Oceanic:



p = .02 (one-sided binom. test)

Step 1: distribution within families

- Sometimes areas (or other possible control factors) do not line up nicely with lower taxa. Semitic:

macrocontinent	mbranch	sbranch	ssbranch	language	ACC
Africa	**West Semitic**	**Southern (West Semitic)**	**Ethiopean**	**Amharic**	**S=A**
Eurasia	**West Semitic**	**Central (West Semitic)**	**South-Central Semitic**	**Arabic (Gulf)**	**S=A**
Eurasia	East Semitic	Akkadian	NA	Akkadian	S=A

- ▶ Posit pseudo-groups, assumed only for the purpose of testing the effect of the macro-area control factor:

  “Eurasian West Semitic” vs. “African West Semitic”

Step 1: distribution within families

- For families that are big enough ($N \geq 5$) we can estimate biases by standard statistical tests (binomial test, χ^2-test etc.)

family.name	majority.response	majority.prop	distribution	macrocontinent	taxonomic.level	family size
Indo-European	S=A	0.74	biased (trend)	Eurasia	stock	65
Muskogean	S=A	1.00	biased (absolute)	Americas	stock	6
Turkic	S=A	1.00	biased (absolute)	Eurasia	stock	8
Dravidian	S=A	1.00	biased (absolute)	Eurasia	stock	8
Uto-Aztecan	S=A	1.00	biased (absolute)	Americas	stock	17
Na-Dene	S=A	1.00	biased (absolute)	Americas	stock	6
Austroasiatic	S=A	1.00	biased (absolute)	Eurasia	stock	8
Salishan	S=A	1.00	biased (absolute)	Americas	stock	5
Mayan	S=A	1.00	biased (absolute)	Americas	stock	5
Benue-Congo	S=A	1.00	biased (absolute)	Africa	stock	16
Tucánoan	S=A	1.00	biased (absolute)	Americas	stock	5
Oceanic	S=A	0.89	biased (trend)	NG-Australia	sbranch	9
Sino-Tibetan	diverse	NA	diverse	Eurasia	stock	45
Pano-Tacanan	diverse	NA	diverse	Americas	stock	5
Nakh-Daghestanian	diverse	NA	diverse	Eurasia	stock	13
Pama-Nyungan	diverse	NA	diverse	NG-Australia	stock	63
Cariban	diverse	NA	diverse	Americas	stock	5

Step 2: extrapolations

- But in addition we have many small uniform families and isolates and information about them is just as important.
- Assumption: *Unknown families are subject to the same developmental trends as known families.*
- Use all information we have about known families to estimate what kinds of trends there are in unknown families:
  - the range of attested values (S=A bias, S≠A bias, diverse), with H_0 probabilities of ⅓ each
  - the actual values in small families
  - the proportion of biased vs. diverse families
- Various techniques for extrapolation (subject of current research by T. Zakharko). One technique:

Step 2: extrapolations

a. Estimate the proportion P_{div} of diverse families among small families on the basis of what we know from large families, using Laplace's *Rule of Succession*:

   $\mathsf{P}(\mathsf{X}_{\mathsf{n}+1} = \text{‘diverse’} \mid \mathsf{k}(\text{‘diverse’})) = \frac{\mathsf{k}+1}{\mathsf{n}+2}$, e.g. if $\frac{1}{8}$ is diverse, assume that $\frac{1+1}{8+2} = .2$ small uniform families are the sole survivors of a diverse family and the rest are the sole survivors of biased families (biased in whatever way)

b. Randomly take P_{div} of small families and declare them ‘diverse’, i.e. think of them as the sole survivors of a diverse larger family

c. For the rest, assume they represent the sole survivors of biased families, with whatever value they have

   Special case: if the small family is diverse (e.g. 1 A and 1B), pick a value at random

Thanks to discussions with Taras Zakharko

Step 2: extrapolations

d. But some of the extrapolated biases may be wrong, as the sole survivor may be exceptional:

- account for the probability of this error by assuming true biases only in proportion to the degree of biases in known families (usually between .8 and .9), here:

	Eurasia	Americas	NG-Australia	Africa
μ(prop.)	0.94	1	0.89	1

and assign the exceptions a randomly chosen alternative value.

Step 2: extrapolations

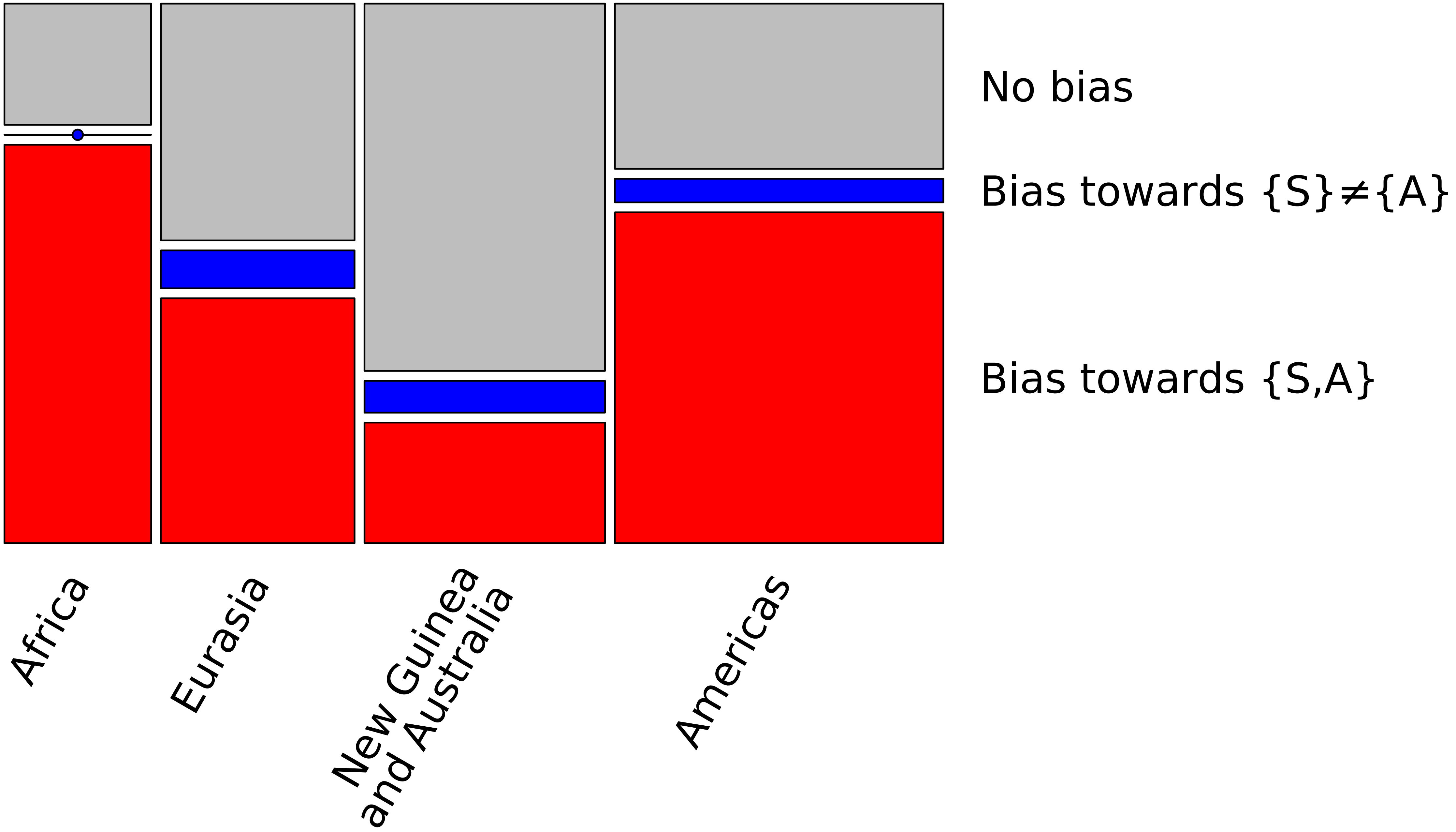
e. Now these extrapolations use random assignments in three places:

- when choosing which small families are ‘diverse’ (we know the proportion, but we don’t know which ones they are)
- when correcting for the possibility that the sole survivors may be exceptions
- when picking a value for small diverse families estimated to be survivors of biased families

These random assignments introduce a statistical error but since they are random, the error can be assumed to be normally distributed

f. Therefore, we can take the mean of a set of random assignments, e.g. the mean of 2,000 extrapolations

Step 3: statistical modeling and testing



"No bias" contains no evidence for or against the hypothesis:
Unless we know the proto-languages, "no bias" can result from either a trend towards S=A or a trend towards S≠A!

Step 3: statistical modeling and testing

- Think of contingency tables as count tables:

macrocontinent	bias	μ(*F*)
Africa	Bias towards {S}≠{A}	0.00
Eurasia	Bias towards {S}≠{A}	2.31
NG-Australia	Bias towards {S}≠{A}	2.59
Americas	Bias towards {S}≠{A}	2.51
Africa	Bias towards {S,A}	19.22
Eurasia	Bias towards {S,A}	15.52
NG-Australia	Bias towards {S,A}	9.54
Americas	Bias towards {S,A}	35.81

- “Loglinear” or “Poisson” models of families under conditions:

$$\left.\begin{array}{l}\log(F) = \alpha + \boldsymbol{\beta_1}M + \beta_2 B + \boldsymbol{\beta_3}M\cdot B \\ \log(F) = \alpha + \boldsymbol{\beta_1}M + \beta_2 B\end{array}\right\} \text{which one fits better?}$$

Step 3: statistical modeling and testing

- Estimate the coefficients following the same basic idea as with the earlier linear model, but using different methods and algorithms (not “least squares” but “maximum likelihood” estimation: find the coefficients so that they best predict the response):
- $\log(\hat{F}) = 2.85 - 2.32B + 0.72Am - .38NGA - .05EUR$
- “S=A” is baseline for B; Africa is baseline for M, so we get
  - $\hat{F} = e^{2.85} = 17.30$ predicted S=A (B=0) in Africa (Am=0, NGA=0, EUR=0)
  - $\hat{F} = e^{2.85-2.31} = 1.71$ predicted S≠A (B=1) in Africa
  - $\hat{F} = e^{2.85-2.31-.38} = 1.17$ predicted S≠A (B=1) in NG-AUS (Am=0, EUR=0)
- Note: in a model without interaction, $\hat{F}$ is exactly the same as what’s expected from the row and colum sums (the expected values “E” in the slides about Pearson residuals in Part I)!

Step 3: statistical modeling and testing

macrocontinent	bias	F	$\hat{F}$ without interaction	$\hat{F}$ with interaction
Africa	Bias towards {S}≠{A}	0.00	1.71	0
Eurasia	Bias towards {S}≠{A}	2.31	1.62	2
NG-Australia	Bias towards {S}≠{A}	2.59	1.17	3
Americas	Bias towards {S}≠{A}	2.51	3.51	3
Africa	Bias towards {S,A}	19.22	17.29	19
Eurasia	Bias towards {S,A}	15.52	16.38	16
NG-Australia	Bias towards {S,A}	9.54	11.83	10
Americas	Bias towards {S,A}	35.81	35.49	36

- With interaction, the model is “saturated”, it (necessarily) fits perfectly (modulo rounding).
- We are interested in the loss of fit when choosing the simpler model without interaction: the “**deviance**” (a.k.a. “likelihood ratio”, “G^2 statistic”) (here similar to the χ^2 value summing the residuals, i.e. the difference between expected and observed)

Step 3: statistical modeling and testing

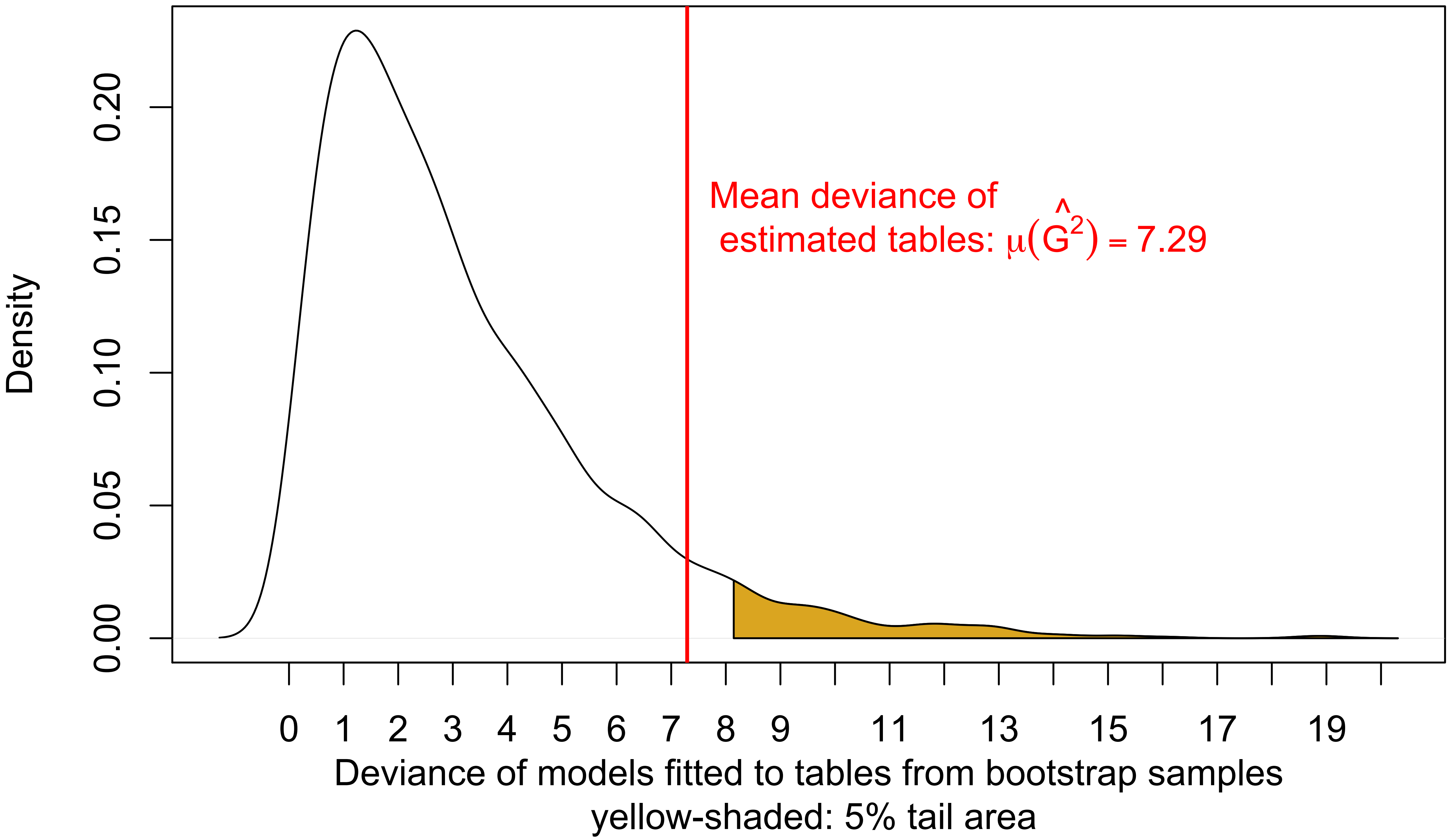
- Test this by simulating what we would observe if the distributions of *F* across levels would be random (“bootstrap sampling”):

  1. randomly assign S≠A or S=A to tables many times, and in each case, compute the deviance

  2. compute the deviance for all 2000 “observed” (i.e. estimated) tables

  3. Check how the mean of the “observed” tables compares to the deviances in the randomly generated tables:

Step 3: statistical modeling and testing

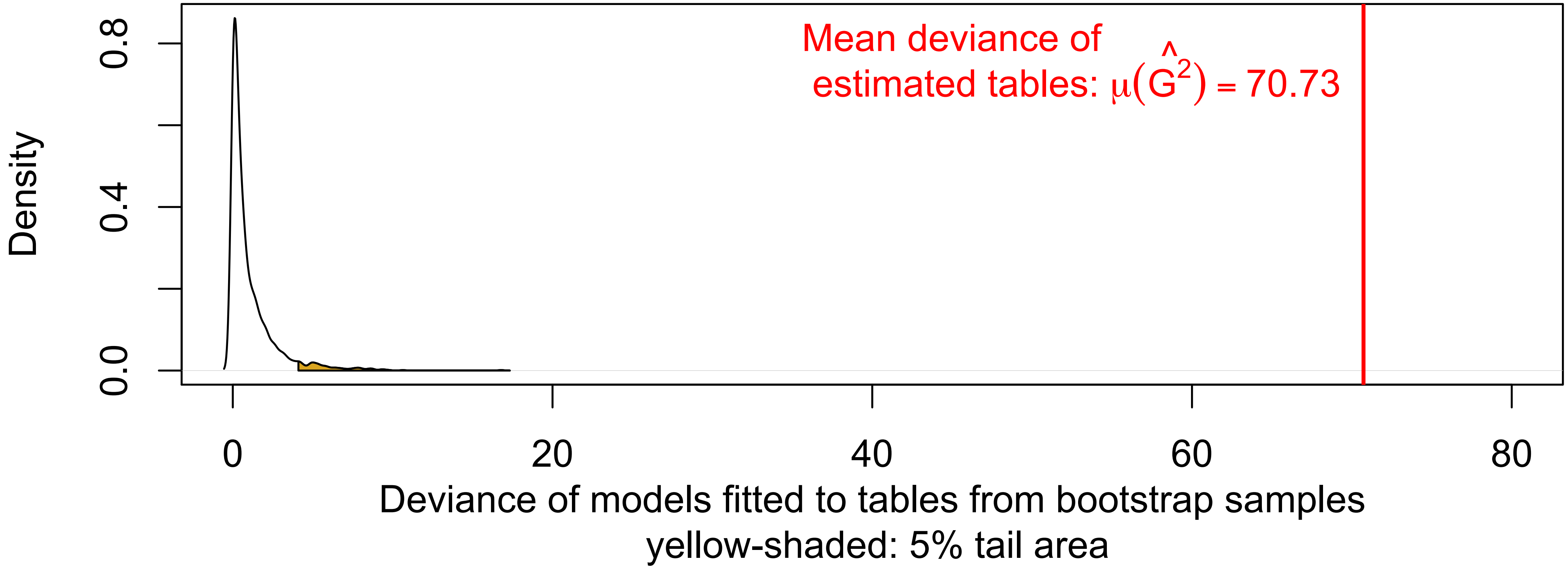


Step 3: statistical modeling and testing

- No evidence for interaction.
- Now ask whether we need the bias factor. Compare:

$$\left.\begin{array}{l} \log(F) = \alpha + \boldsymbol{\beta_1} M + \beta_2 B \\ \log(F) = \alpha + \boldsymbol{\beta_1} M \end{array}\right\} \text{which one fits better?}$$

- Compute deviance between the models as before:



Step 3: statistical modeling and testing

Comparing the fits:

macrocontinent	bias	*F*	with bias factor	area only
Africa	Bias towards{S}≠{A}	0.00	1.71	9.50
Eurasia	Bias towards{S}≠{A}	2.31	1.62	9.00
NG-Australia	Bias towards{S}≠{A}	2.59	1.17	6.50
Americas	Bias towards{S}≠{A}	2.51	3.51	19.50
Africa	Bias towards{S,A}	19.22	17.29	9.50
Eurasia	Bias towards{S,A}	15.52	16.38	9.00
NG-Australia	Bias towards{S,A}	9.54	11.83	6.50
Americas	Bias towards{S,A}	35.81	35.49	19.50

Conclusion: the smallest model that still fits well includes BIAS as a factor. This supports the hypothesis.

Interim summary

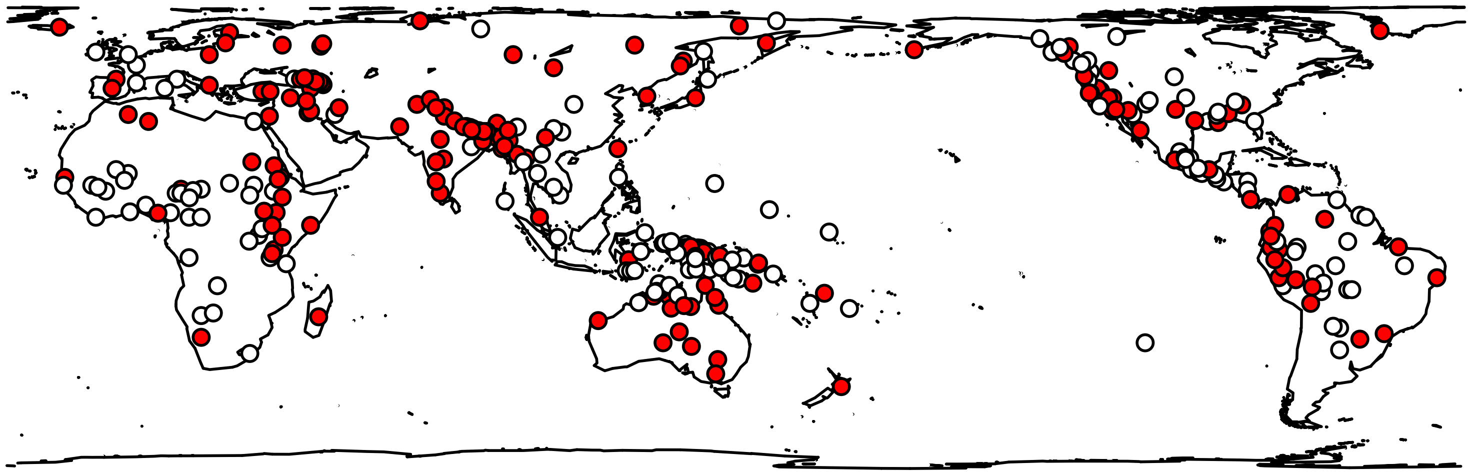
- Three steps:
  - Evaluate biases within families given all factors of interest
  - Extrapolate bias estimates to small families and singletons, using all information available
  - Model the distribution of biases given the factors of interest
- Because we use generalized linear modeling, this scales to any complexity of the hypothesis
- Predictors can be anything!
- One more case study with a possible interaction between two structural variables: case and word order

Another case study: the distribution of case over word order

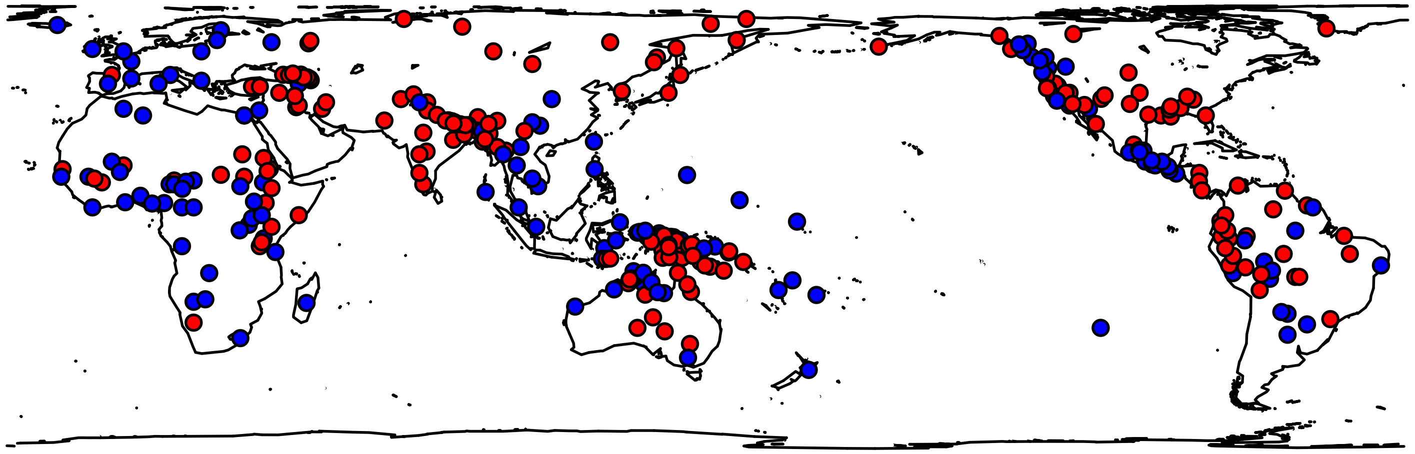
- Hawkins 2004: Verb-final languages favor rich case “for reasons of on-line efficiency” (‘rich’ = distinct coding of agent and patient)

- Nichols 1992, Siewierska 1996, Dryer 1989, 2000, 2005, Bickel & Nichols 2006: the distribution of both case and word order is heavily affected by areal patterns:

Rich case

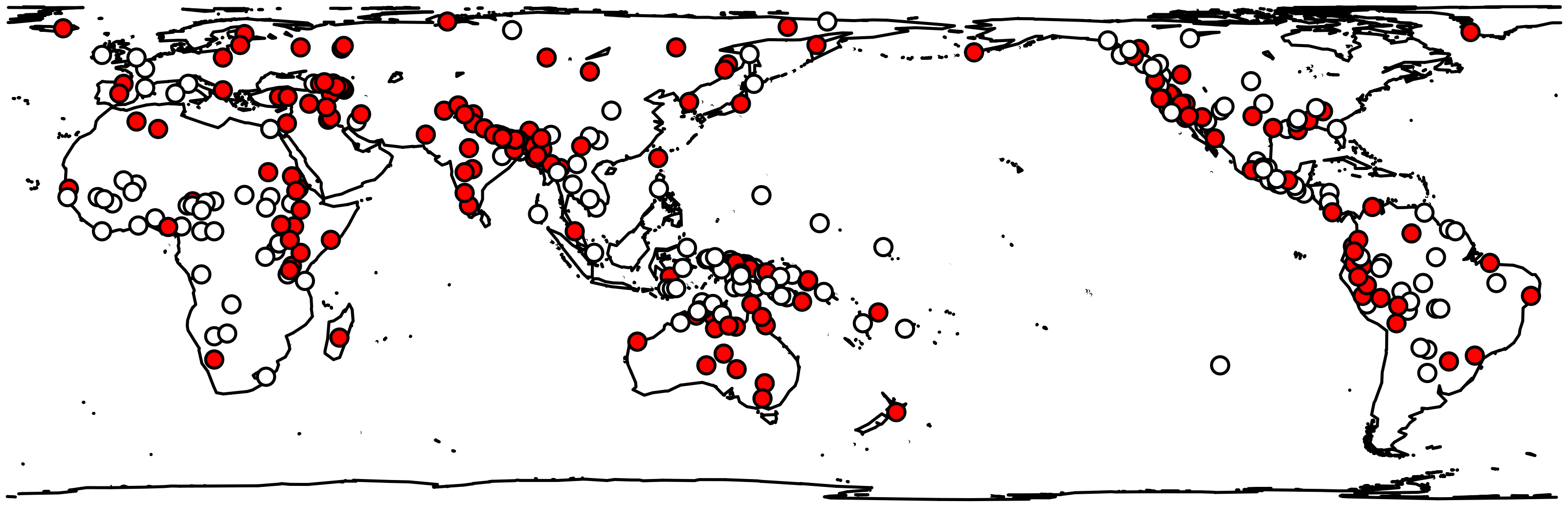


OV vs VO order



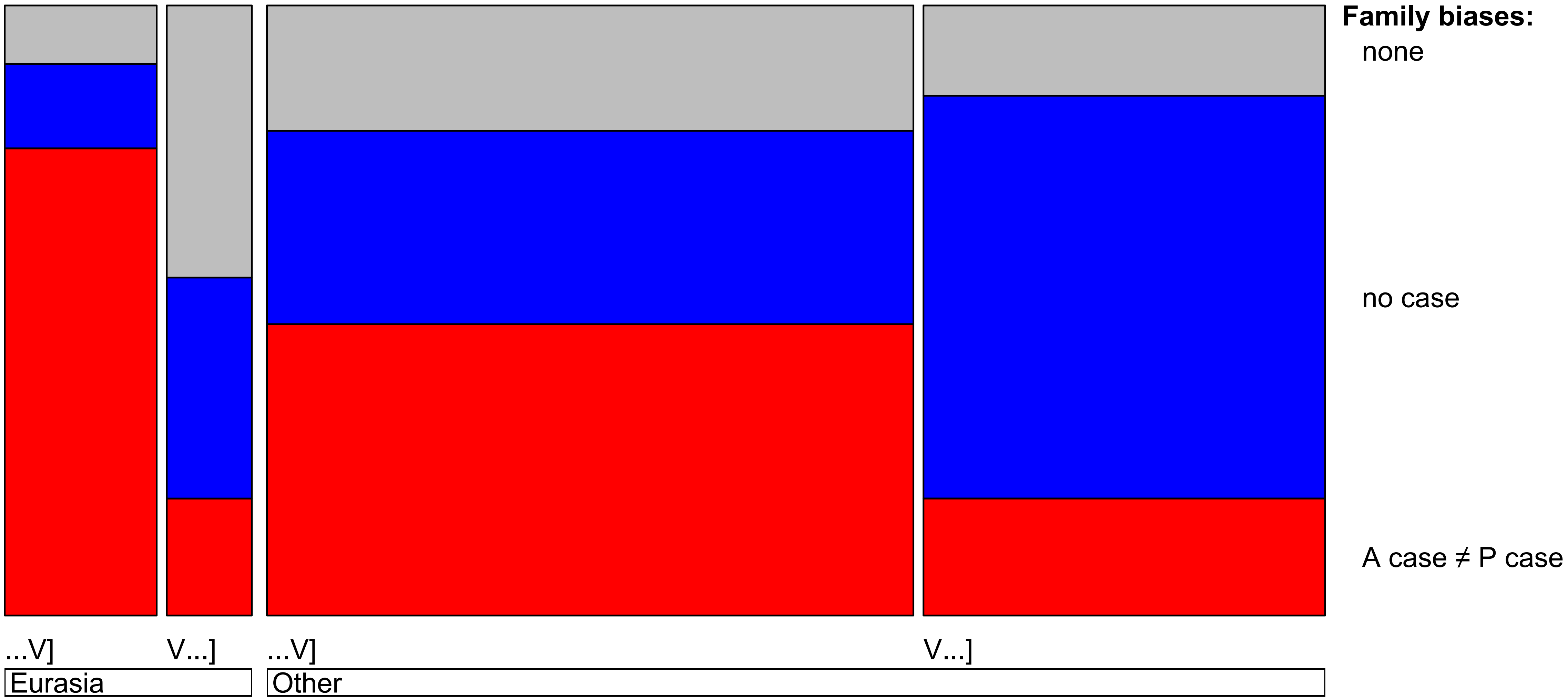
Case Study: the distribution of case over word order

- Data on rich case from Comrie 2005 (WALS) and AUTOTYP, 1% mismatches
- Data on word order from Dryer 2005 (WALS) and AUTOTYP, 0% mismatches
- Total datapoints with information on both variables: $N = 330$
- Stocks with more than one member: $N = 51$
- Various areal confounding factors, at different levels of resolution
- Focus here on Eurasia:



Case Study: the distribution of case over word order

- Estimate biases in large families, conditioned on ±EURASIA and ±VO
- Extrapolate to ~~small families and~~ isolates
- Compute the mean table:



Case Study: the distribution of case over word order

bias	word order	area	F
neutral	final	Eurasia	3.46
non-neutral	final	Eurasia	19.14
neutral	non_final	Eurasia	5.07
non-neutral	non_final	Eurasia	2.68
neutral	final	Other	32.65
non-neutral	final	Other	49.19
neutral	non_final	Other	42.26
non-neutral	non_final	Other	12.27

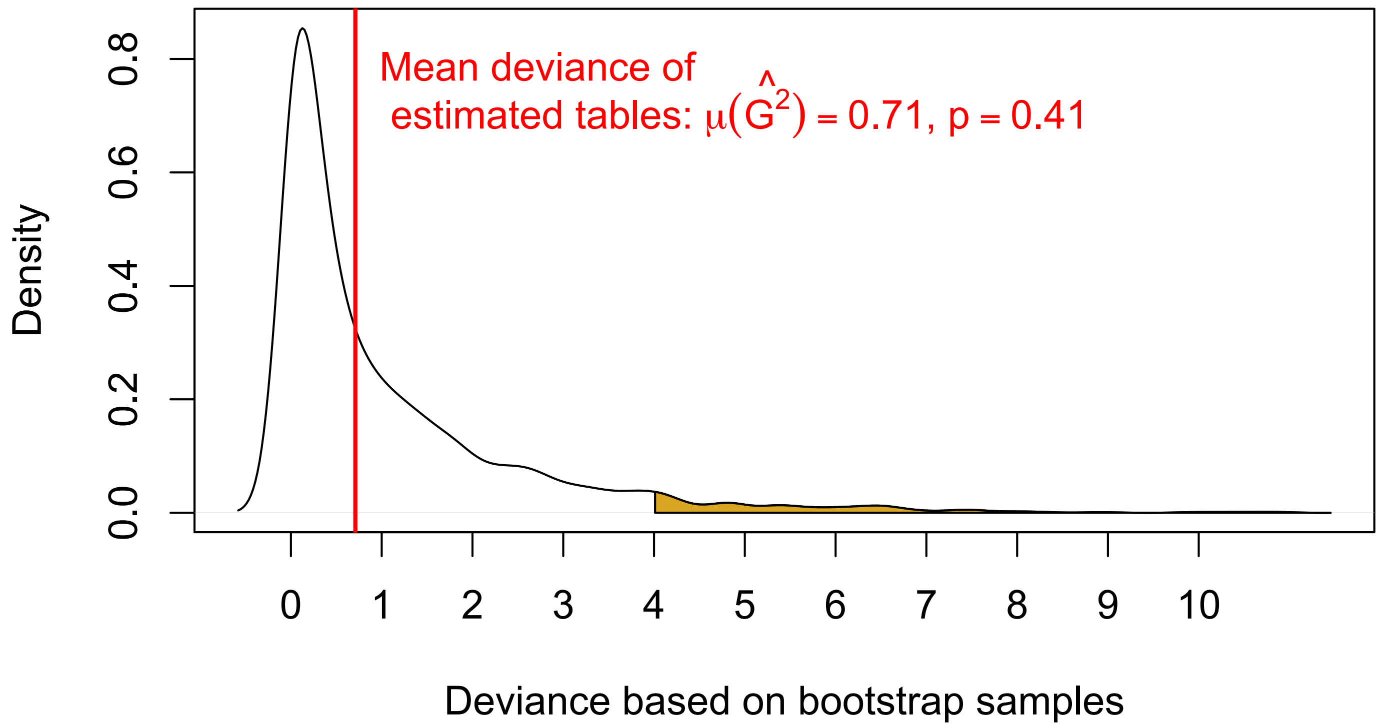
bias	word order	area	fitted
neutral	final	Eurasia	3.69
non-neutral	final	Eurasia	18.31
neutral	non_final	Eurasia	4.31
non-neutral	non_final	Eurasia	3.69
neutral	final	Other	32.31
non-neutral	final	Other	49.69
neutral	non_final	Other	42.69
non-neutral	non_final	Other	11.31

- Compare models (A: area; W: word order, B: case bias)

  $$\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + \beta_4 AW + \beta_5 BA + \beta_6 BW \mathbf{+ \beta_7 BAW}$$

  $$\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + + \beta_4 AW + \beta_5 BA + \beta_6 BW$$

  - the simpler model fits just as well:

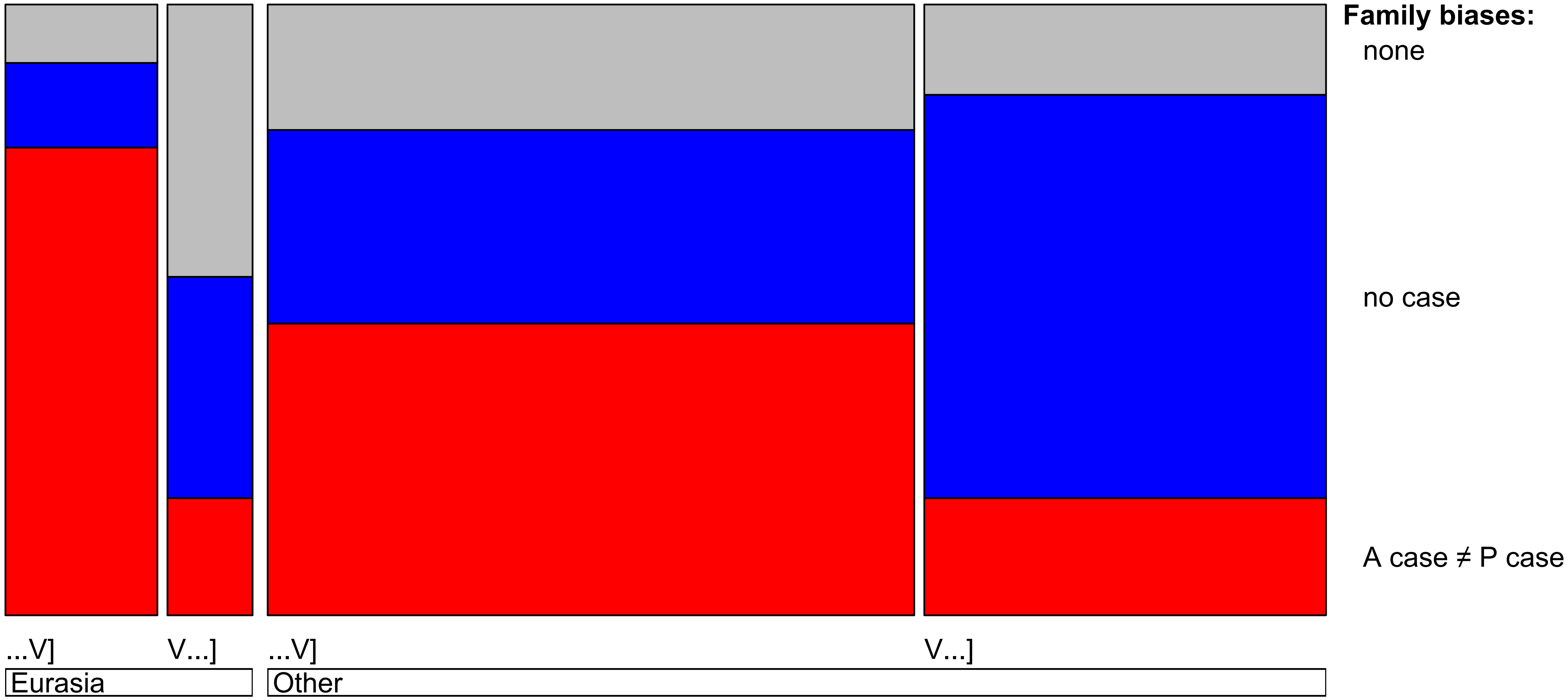


Case Study: the distribution of case over word order

- $\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + + \beta_4 AW + \beta_5 BA + \beta_6 BW$

  means that the bias interacts with area and with word order and the area interacts with word order

- How good is the evidence for the interaction of the bias with area ($\beta_5 BA$) and with word order ($\beta_6 BW$)?

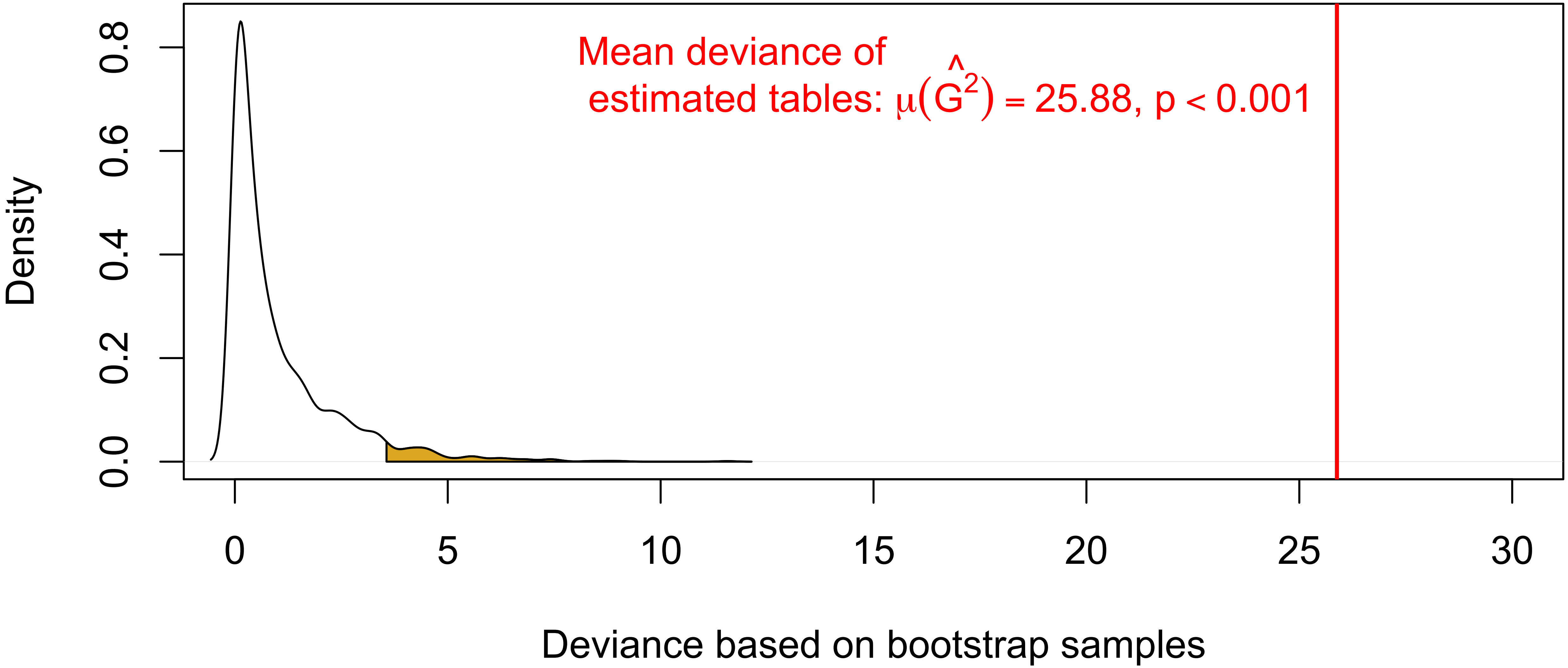


Case Study: the distribution of case over word order

- Again, compare models:

$$\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + \beta_4 AW + \beta_5 BA + \boldsymbol{\beta_6 BW}$$

$$\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + \beta_4 AW + \beta_5 BA$$



Case Study: the distribution of case over word order

- Without word order, we can’t predict the distribution well:

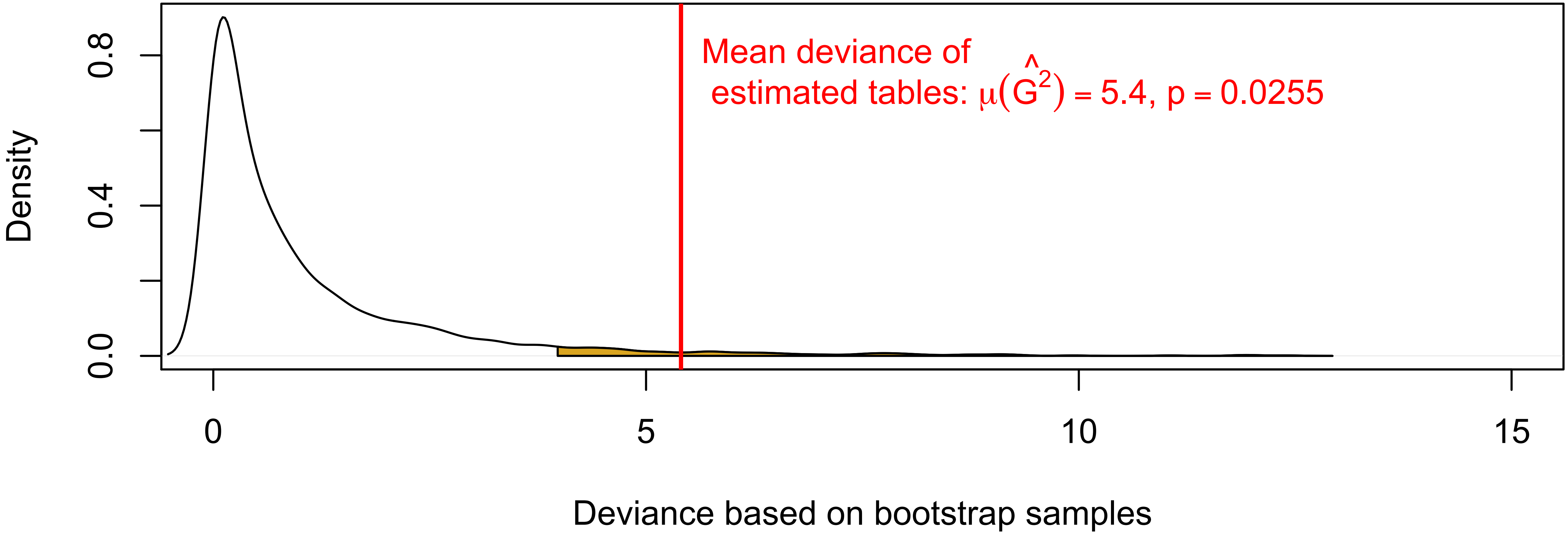
bias	word order	area	F	with β6BW	without β6BW
neutral	final	Eurasia	3.46	3.69	5.87
non-neutral	final	Eurasia	19.14	18.31	16.13
neutral	non_final	Eurasia	5.07	4.31	2.13
non-neutral	non_final	Eurasia	2.68	3.69	5.87
neutral	final	Other	32.65	32.31	45.22
non-neutral	final	Other	49.19	49.69	36.78
neutral	non_final	Other	42.26	42.69	29.78
non-neutral	non_final	Other	12.27	11.31	24.22

Case Study: the distribution of case over word order

- And what about the area effect?

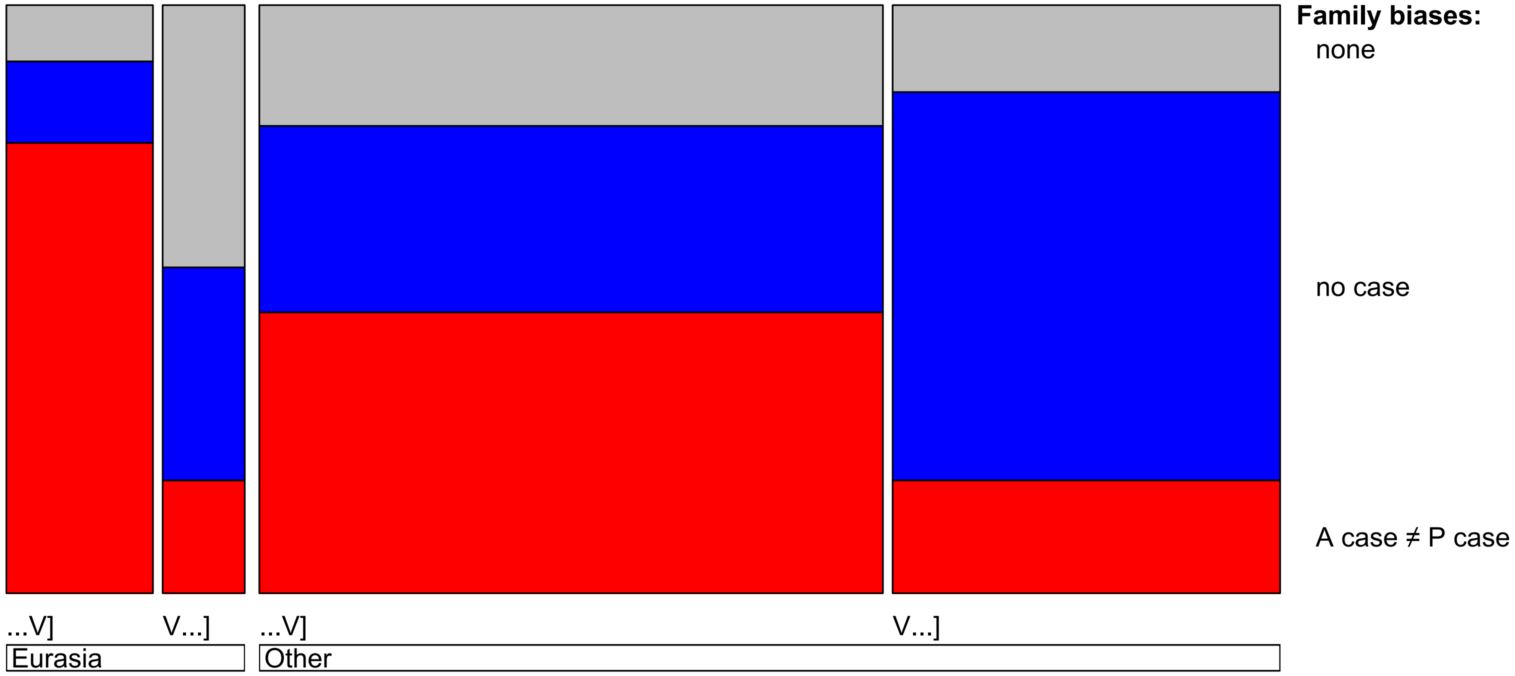
  $\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + \beta_4 AW + \beta_5 BW + \boldsymbol{\beta_6 BA}$

  $\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + \beta_4 AW + \beta_5 BW$



Case Study: the distribution of case over word order

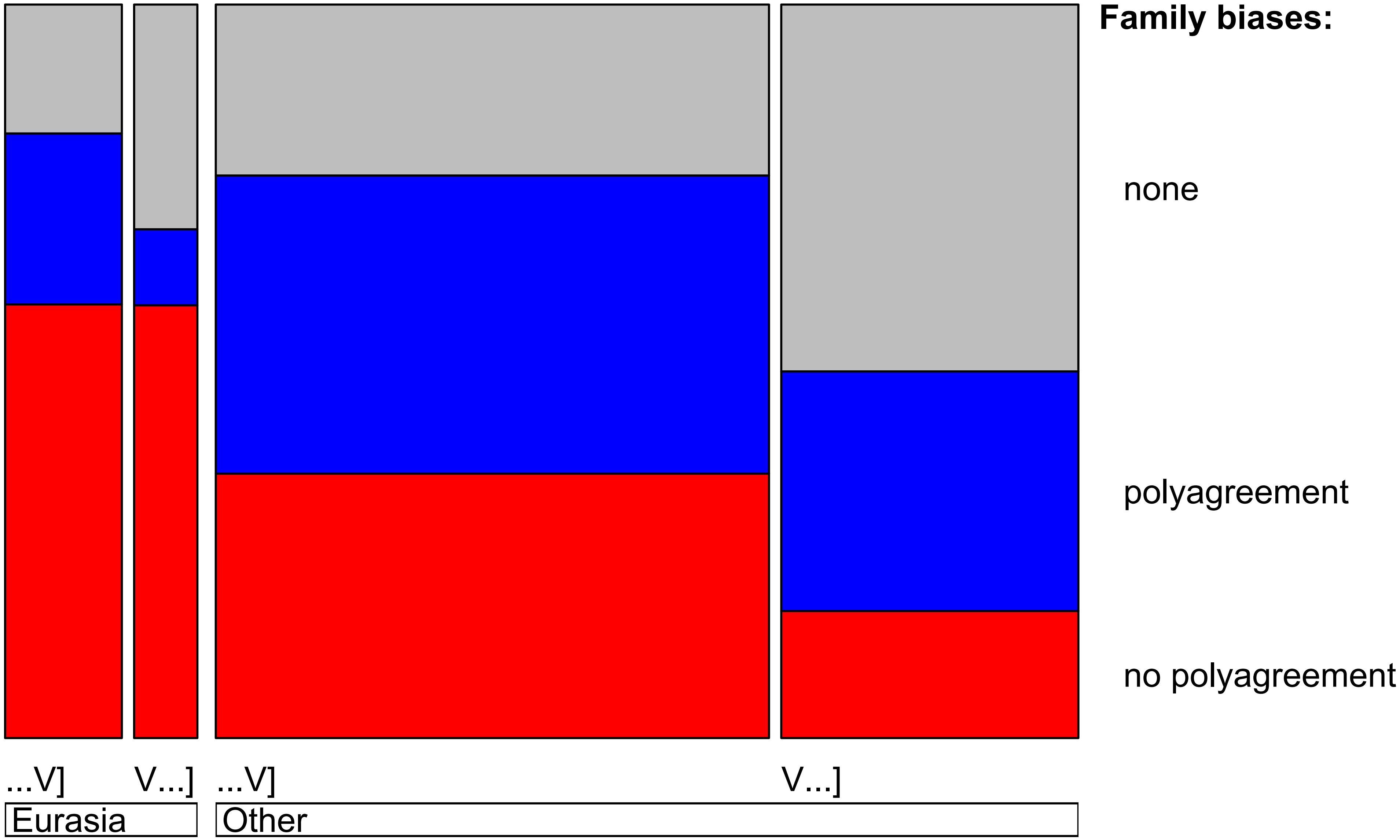
- Conclusions on case study:
- in order to predict the frequencies of biases reasonably well, we need to know
  - the relative frequencies across word order
  - the relative frequencies across areas



- In other words, the distribution of biases towards or against case depends on both area (B·A) and word order (B·W), but these dependencies do not interact (B·A·W)

Case Study: the distribution of agreement over word order

- Hawkins 2004 also hypothesizes that V-final languages disfavor rich agreement, i.e. agreement with both A and P arguments (because it's not 'needed').
- Mean extrapolated tables (extrapolation to isolates):



Case Study: the distribution of agreement over word order

- B·A·W: *p* = .12, so we can again simplify to
- $\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + \beta_4 AW + \beta_5 BA + \beta_6 BW$
  - But now $\beta_6 BW$ doesn’t contribute either, p = .55
  - While $\beta_5 BA$ still does, p = .001

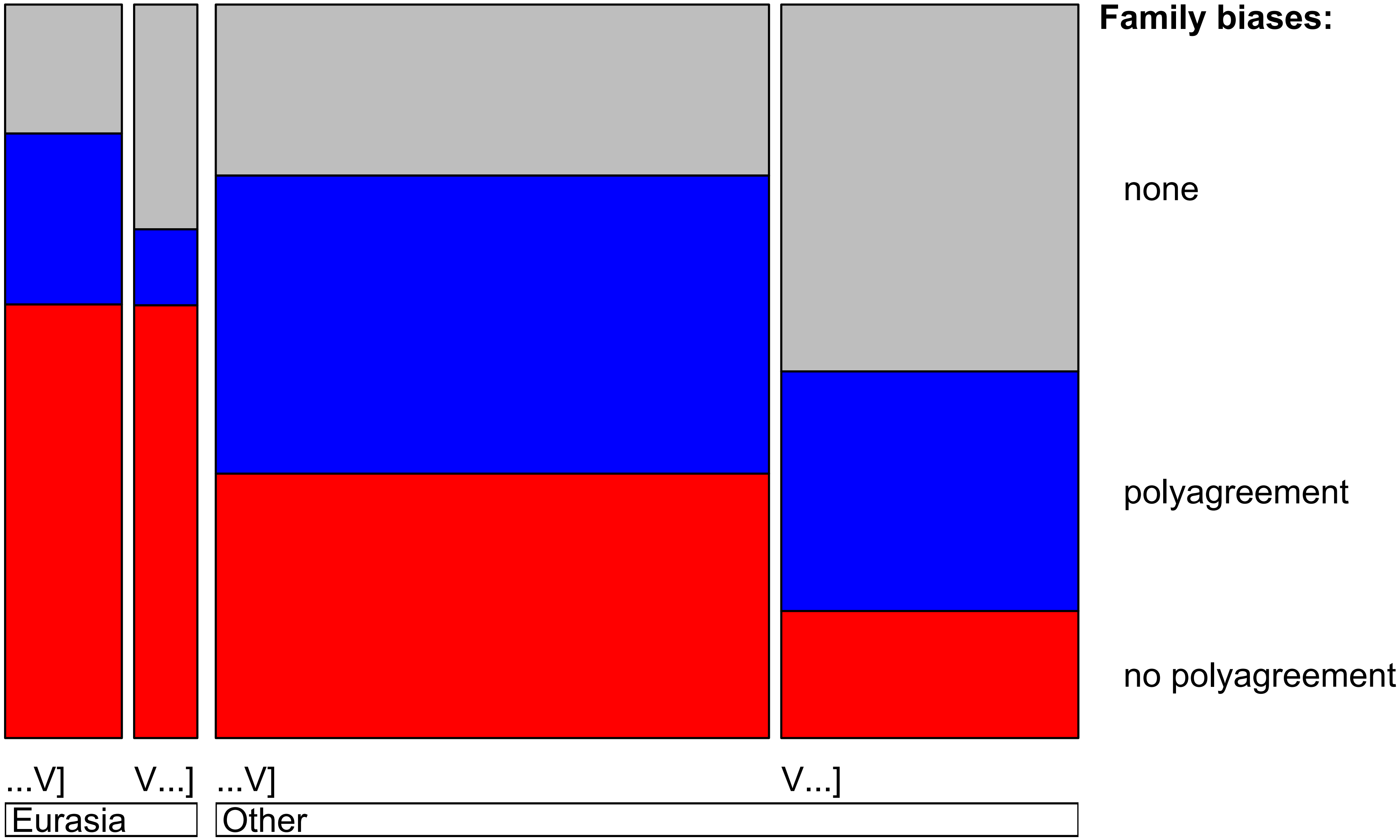
Case Study: the distribution of agreement over word order

- Best-fitting model: $\log(F) = \alpha + \beta_1 B + \beta_2 A + \beta_3 W + \beta_4 AW + \beta_5 BA$

bias (B)	word order (W)	area (A)	F	with βBA and βBW	with BA only	with BW only
POLYAGR	final	Eurasia	5.59	4.55	4.83	9.71
without	final	Eurasia	14.19	15.45	15.17	10.29
POLYAGR	non_final	Eurasia	1.35	2.45	2.17	4.74
without	non_final	Eurasia	7.67	6.55	6.83	4.26
POLYAGR	final	Other	43.88	45.45	46.69	40.29
without	final	Other	38.93	37.55	36.31	42.71
POLYAGR	non_final	Other	18.95	17.55	16.31	15.26
without	non_final	Other	10.04	11.45	12.69	13.74

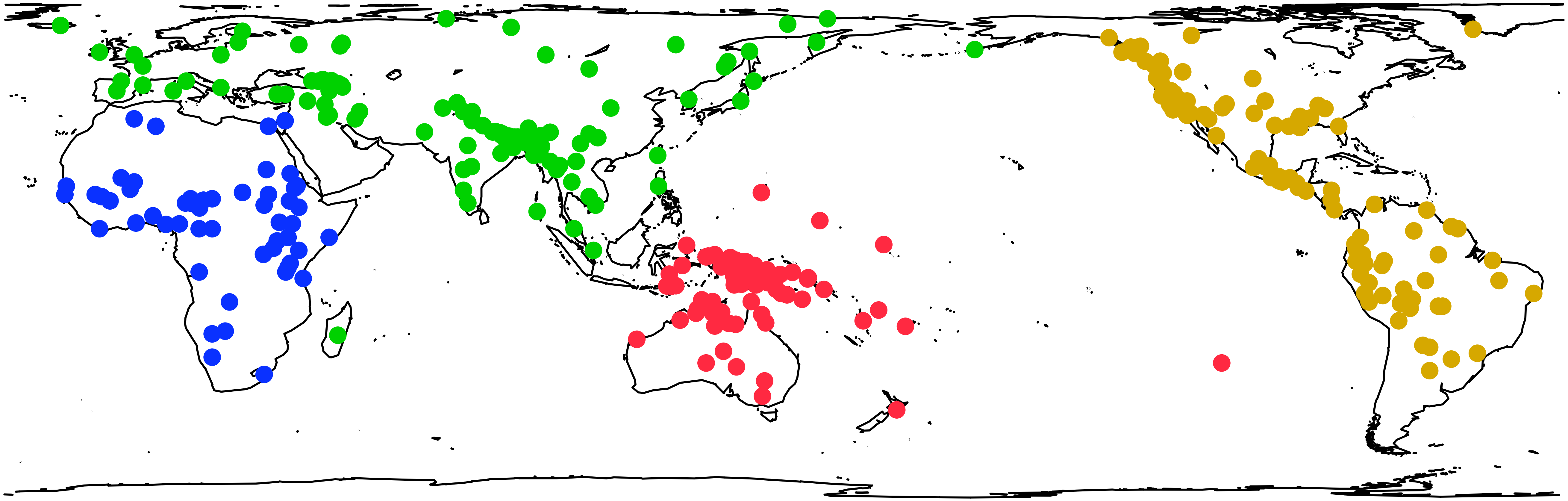
Case Study: the distribution of agreement over word order

- Conclusion: no evidence for non-V-final languages to universally disfavor polyagreement; the trend is too small to be significant.
- But clear evidence for a significant decrease of polyagreement in Eurasia



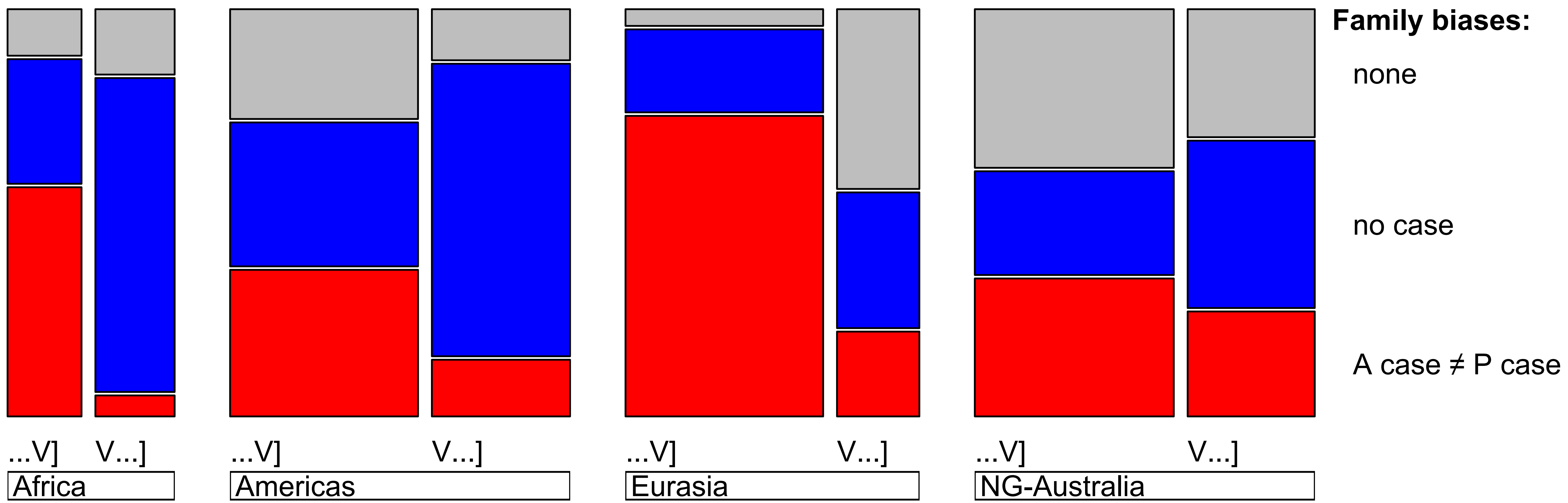
Wouldn’t it be simpler to test effects within each area?

- Dryer (1989, 2000) proposes to look at distributions separately, within each area and then count in how many areas we get an effect
- For example with “macro-continents”:



Wouldn’t it be simpler to test effects within each area?

- Mean of extrapolated tables:



- Results from testing the B·W (bias * word order) interaction within each macrocontinent separately:

p=.001 p=.021 p=.024 p=.234

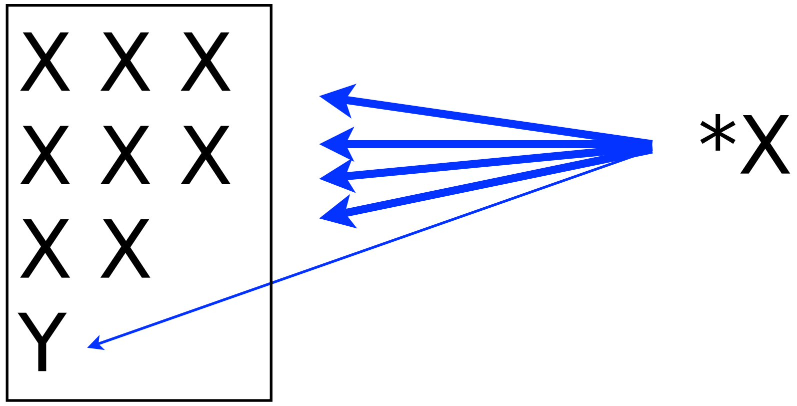
- Compare with overall effect, $p < .001$

Overall conclusions

- If we find a universally significant trend in family biases — either on its own (example: anti-ergative bias) or conditioned by a structural factor (example: case in V-final groups) — this is evidence for a universal
- because such trend reflects developments in each family under two scenarios
  - the proto-language went against the bias and then daughter languages moved away from this
  - the proto-language showed the bias alreadt and then daughter languages kept this
  - because π(keep the bias) > π(loose the bias)
- This interpretation is correct to the extent that $\pi \gg 0$.

An alternative method: g-sampling

- Dryer (1998, 2000): if the members of a family have all the same type, this is because they inherited it from the proto-language:



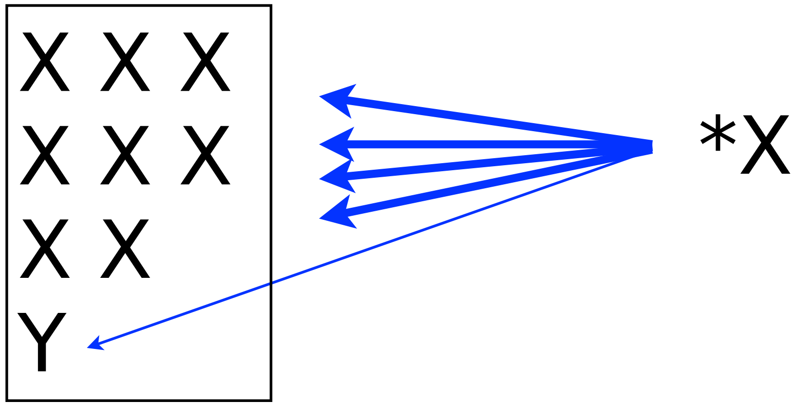
X X X
X X X
X X
Y

*X

- If a family has many members in a database and most of them have the same type X, inherited from the proto-language, we shouldn't count all Xs when testing universals

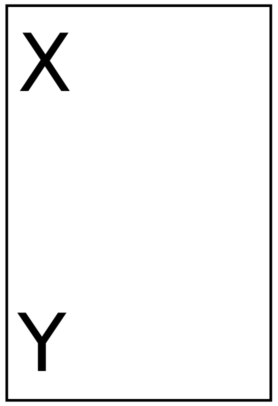
An alternative method: g-sampling

- Therefore, we should remove the “replicated” Xs



X X X
X X X
X X
Y

*X



X

Y

- This is “genealogically balanced” sampling or “g-sampling”: each family contributes only one type

An alternative method: g-sampling

stock	Language	VP	REL
Adamawa-Ubangi	Day	VO	NRel
Adamawa-Ubangi	Doyayo	VO	NRel
Adamawa-Ubangi	Gbeya Bossangoa	VO	NRel
Adamawa-Ubangi	Linda	VO	NRel
Adamawa-Ubangi	Mbodomo	VO	NRel
Adamawa-Ubangi	Mbum	VO	NRel
Adamawa-Ubangi	Mondunga	VO	NRel
Adamawa-Ubangi	Mumuye	VO	NRel
Adamawa-Ubangi	Nzakara	VO	NRel
Adamawa-Ubangi	Samba Leko	VO	NRel
Adamawa-Ubangi	Sango	VO	NRel

→ 1 g-unit = 1 datapoint

An alternative method: g-sampling

- Things get complicated again when we families are split by conditions:
- In this case, we look at the trend in the response variable (e.g. position of relative clauses):
  - when we find a bias within a condition, we interpret this as a reflex of the proto-language and reduce the cases to 1
  - what deviates from the our assumed proto-language stays

An alternative method: g-sampling

stock	Language	VP	REL
Sino-Tibetan	Bai	VO	RelN
Sino-Tibetan	Cantonese	VO	RelN
Sino-Tibetan	Hakka	VO	RelN
Sino-Tibetan	Mandarin	VO	RelN
Sino-Tibetan	Karen (Bwe)	VO	NRel
Sino-Tibetan	Karen (Pwo)	VO	NRel
Sino-Tibetan	Karen (Sgaw)	VO	NRel
Sino-Tibetan	Kayah Li (Eastern)	VO	NRel
Sino-Tibetan	Achang	OV	RelN
Sino-Tibetan	Akha	OV	RelN
Sino-Tibetan	Apatani	OV	RelN
Sino-Tibetan	Athpare	OV	RelN
Sino-Tibetan	Balti	OV	RelN
Sino-Tibetan	Burmese	OV	RelN
Sino-Tibetan	Byangsi	OV	RelN
Sino-Tibetan	Camling	OV	RelN
Sino-Tibetan	Chantyal	OV	RelN
Sino-Tibetan	Chepang	OV	RelN
Sino-Tibetan	Chin (Siyin)	OV	RelN
Sino-Tibetan	Mishmi (Digaro)	OV	RelN
Sino-Tibetan	Dimasa	OV	RelN
Sino-Tibetan	Gallong	OV	RelN
Sino-Tibetan	Gurung	OV	RelN
Sino-Tibetan	Hani	OV	RelN
Sino-Tibetan	Hayu	OV	RelN
Sino-Tibetan	Jinghpo	OV	RelN
Sino-Tibetan	Khaling	OV	RelN
Sino-Tibetan	Kham	OV	RelN
Sino-Tibetan	Lahu	OV	RelN
Sino-Tibetan	Limbu	OV	RelN
Sino-Tibetan	Maru	OV	RelN
Sino-Tibetan	Meithei (Manipuri)	OV	RelN
Sino-Tibetan	Mising	OV	RelN
Sino-Tibetan	Mao Naga	OV	RelN
Sino-Tibetan	Nar-Phu	OV	RelN
Sino-Tibetan	Newar (Dolakha)	OV	RelN
Sino-Tibetan	Newar (Kathmandu)	OV	RelN
Sino-Tibetan	Nocte	OV	RelN
Sino-Tibetan	Purki	OV	RelN
Sino-Tibetan	Rawang	OV	RelN
Sino-Tibetan	Sikkimese	OV	RelN
Sino-Tibetan	Tamang	OV	RelN
Sino-Tibetan	Thulung	OV	RelN
Sino-Tibetan	Tibetan (Modern Literary)	OV	RelN
Sino-Tibetan	Angami Naga	OV	NRel
Sino-Tibetan	Garo	OV	NRel
Sino-Tibetan	Pattani	OV	NRel

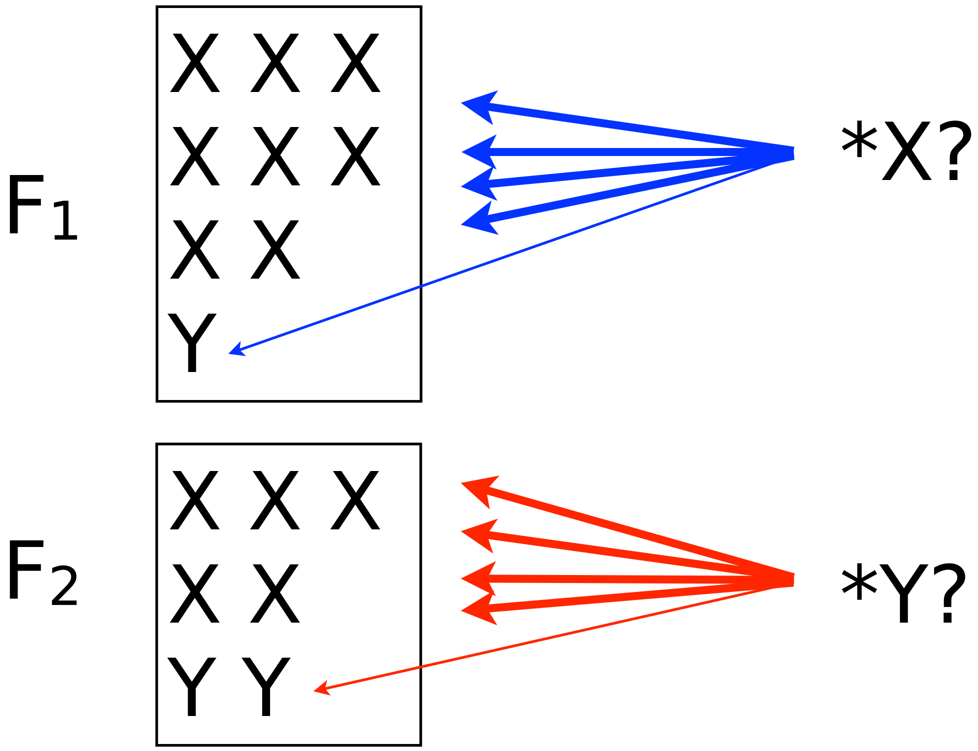
G-units:

VO	**RelN**	4	1	(overall trend in ST)	1 unit
VO	**NRel**	4	4	(deviant within VO)	1 unit
OV	**RelN**	36	1	(overall trend in ST)	1 unit
OV	**NRel**	3	3	(deviant within OV)	3 units

Bickel 2008, including algorithm

An alternative method: g-sampling

Problem: this method (just like most recent studies on “stability indices”) misses the possibility that a bias cannot only arise from faithful inheritance from the proto-language but also from the effects of a universal that favors retention of a type (Maslova 2000):



F_1

X X X
X X X
X X
Y

*X?

F_2

X X X
X X
Y Y

*Y?

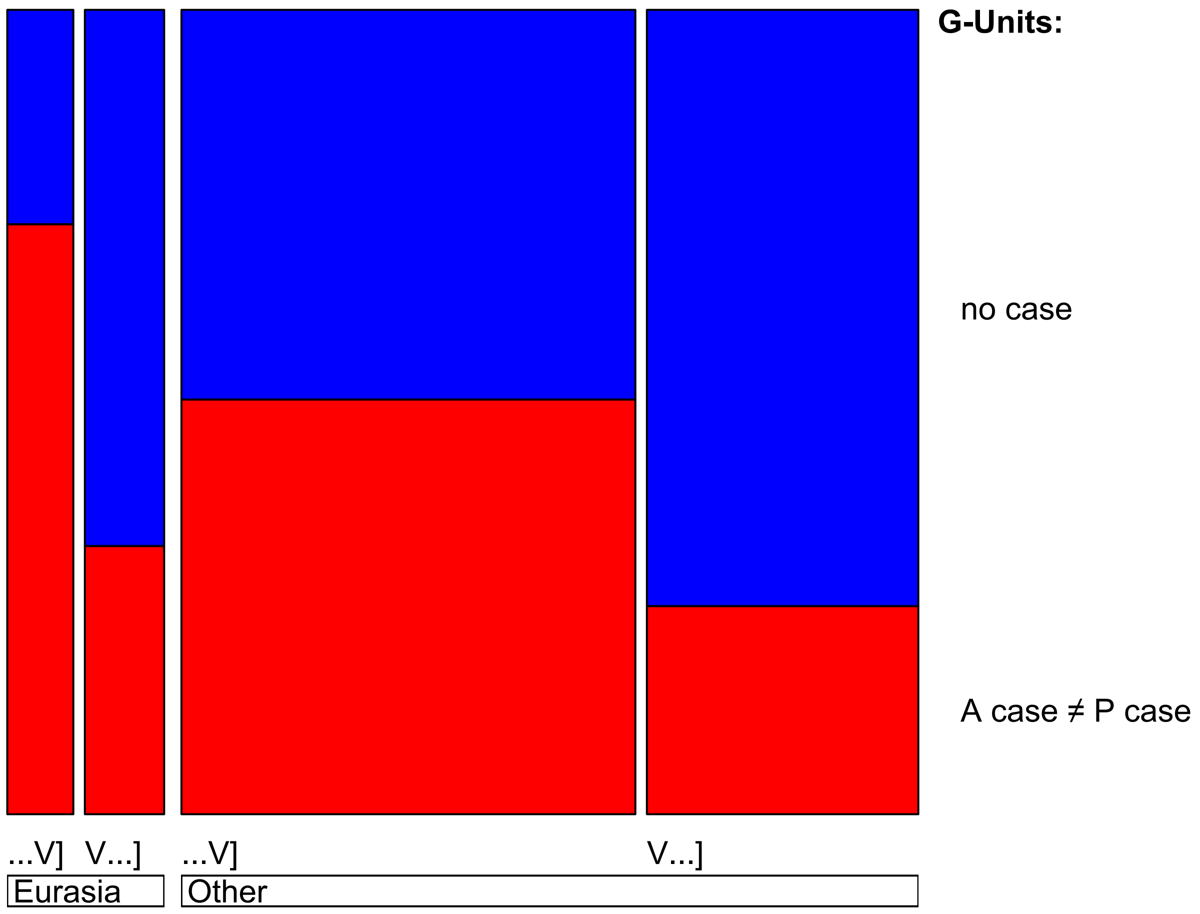
$$\pi(Y \succ X) > \pi(X \succ Y)$$

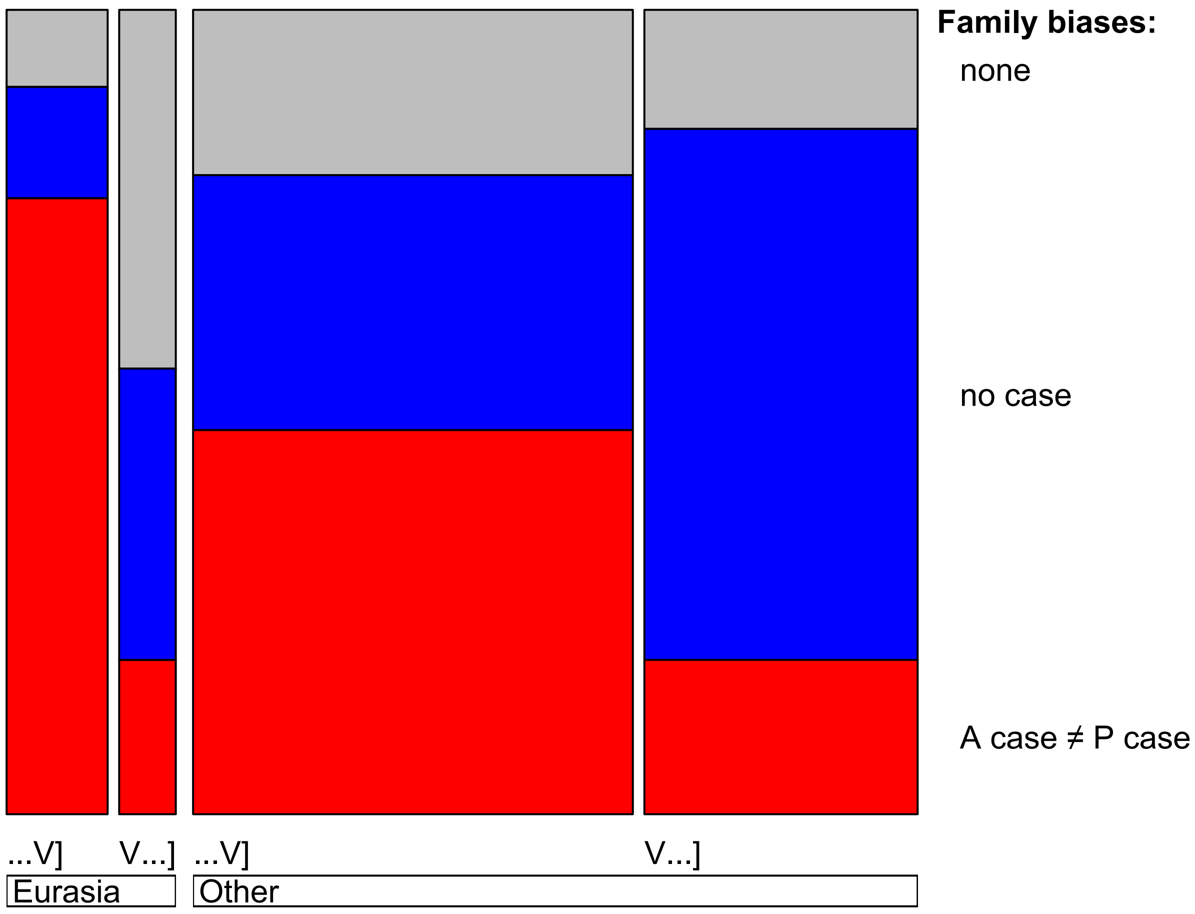
There is no reason why we should only look at hypothesized evidence from change (Y≻X) and not also from retention (X)!

An alternative method: g-sampling

- However, g-sampling has a very useful, practical cousin: pre-defined standard samples, e.g. the WALS 200-languages sample.
- How do the methods compare in practice?
- Example: case ~ word order study

G-Sampling vs. Family Bias Method





	g-sampling	family biases
B·A·W	0.480	0.410
B·A	0.115	0.026
B·W	0.001	0.001

G-Sampling vs. Family Bias Method

- Conclusion: g-sampling picks up some but not all signals.
- Useful for pilot study before collecting large dataset as needed for the family-bias method

A final note on family biases/g-sampling: within-language variation

- In the past, typologist often concentrated on “per language” data, requiring at most one datapoint per language. (= the standard in the *World Atlas of Language Structure*)
- But for many variables and many languages, we find multiple values per language
  - Classical response: reduce the diversity before measuring it (again!)
  - for example
  - take the mean of S=A structures:

Within-language variation

Hindi (Bickel & Yādava 2000)

a. *Ravī ā-yā.*
R.(M)[NOM] come-PST.PTCP.3sM
‘Ravi came.’

b. *Ravī=ne roṭī khā-yī.*
R.(M)=ERG bread(sF)[NOM] eat-PST.PTCPsF
‘Ravi ate some roti.’

c. *Ravī=ne roṭī=ko khā-yā.*
R.(M)=ERG bread(sF)=DAT eat-PST.PTCPsM
‘Ravi ate the roti.’

d. *Ravī roṭī khā-e-gā.*
R.(M)[NOM] bread(sF)[NOM] eat-3s-FUTsM
‘Ravi will eat some bread.’

e. *tumhārā /*tum kal nahī̃ ā-kar kuc kām nahī̃ ho-gā.*
2mhGEN 2mhNOM tomorrow not come-CVB some work(sM)[NOM] not be.3s-FUTsM
‘If you don’t come tomorrow, there won’t be much work.’

f. *tumhārā /*tum phāṛ-e yah gā̃ṭh nahī̃ phaṭ-e-gī.*
2mhGEN 2mhNOM split-CVB DEM log(sF)[NOM] not become.split-3s-FUTsF
‘Even if you (try to) split it, this log won’t split apart.’

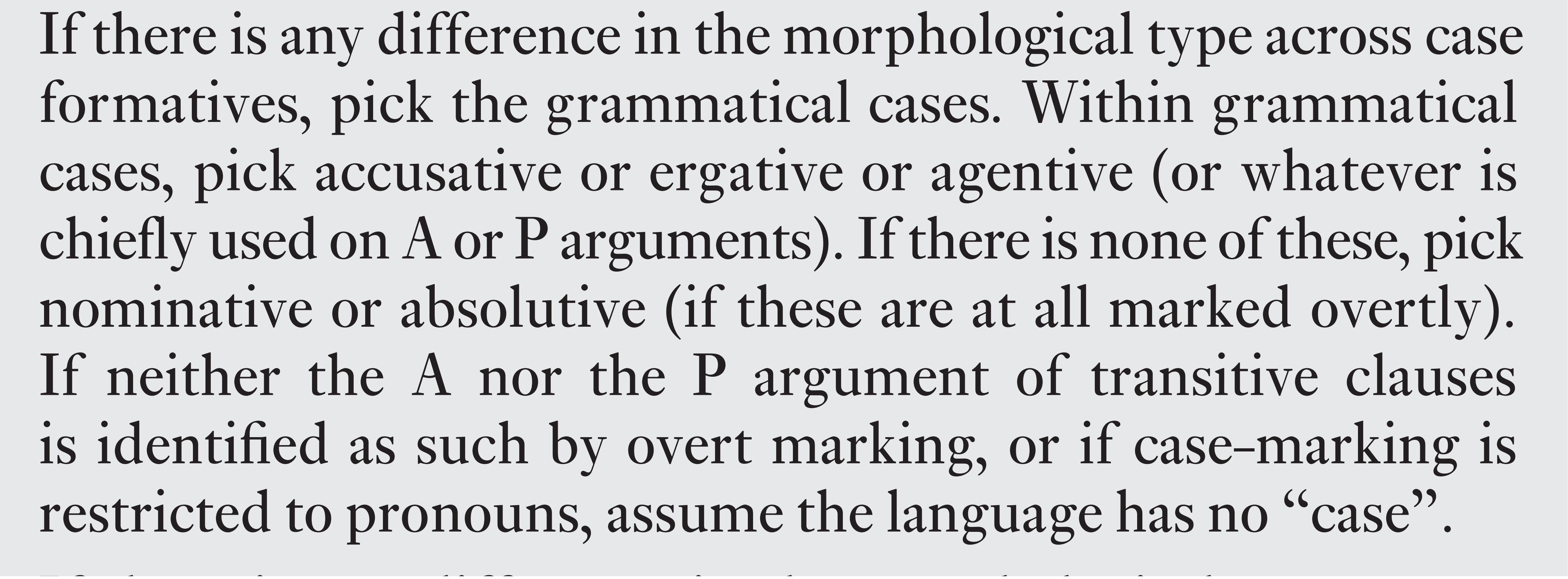
Within-language variation

Hindi: $\mu(S=A) = 0.70$

Verb form	Reference	Clause type	Alignment
PTCP_based	N-high	finite	S≠A≠O
non-PTCP_based	N-high	nonfinite	S=A≠O
non-PTCP_based	N-high	finite	S=A≠O
PTCP_based	N-high	nonfinite	S=A≠O
PTCP_based	N-low	finite	S=O≠A
non-PTCP_based	N-low	finite	S=A=O
PTCP_based	Pro	nonfinite	S=A≠O
PTCP_based	Pro	finite	S≠A≠O
non-PTCP_based	Pro	finite	S=A≠O
non-PTCP_based	Pro	nonfinite	S=A≠O

Within-language variation

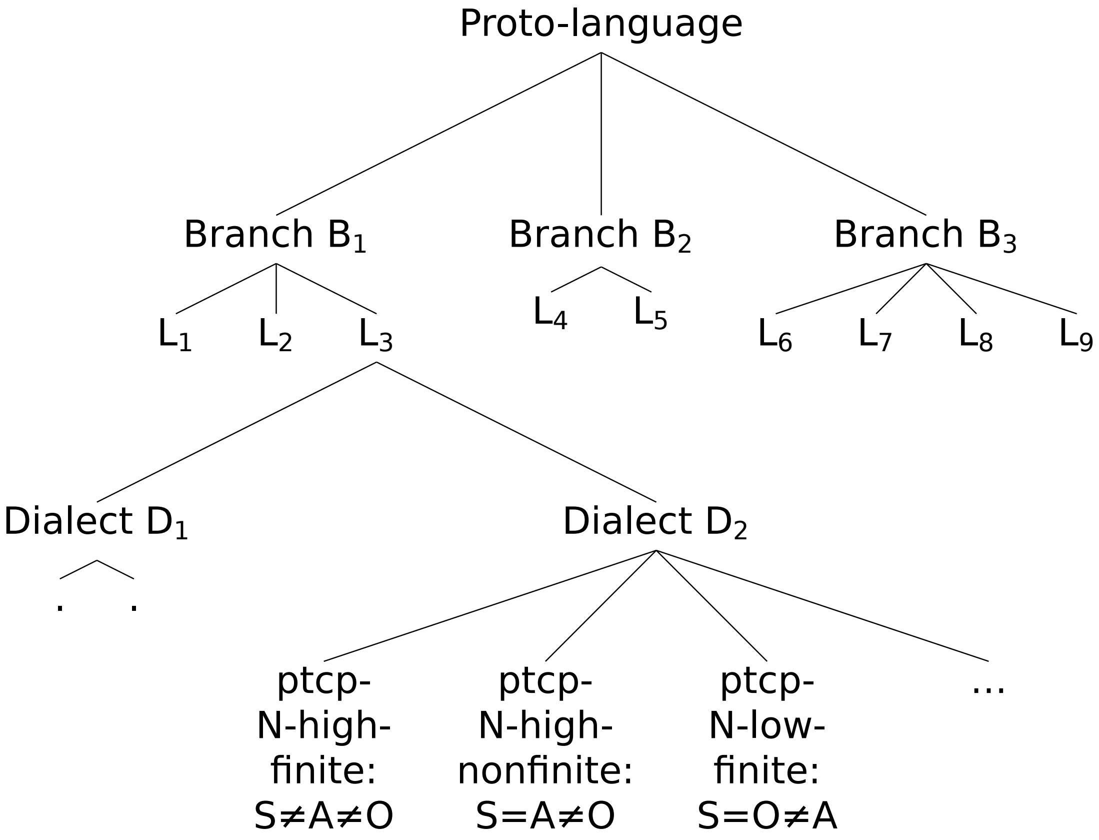
- or take only “basic” structures (“exemplars” in Bickel & Nichols 2002, 2005):
  - basic word order (Dryer 2005): main clause, declarative, pragmatically neutral, lexical arguments, most frequent
  - case exemplar (Bickel & Nichols 2005):



If there is any difference in the morphological type across case formatives, pick the grammatical cases. Within grammatical cases, pick accusative or ergative or agentive (or whatever is chiefly used on A or P arguments). If there is none of these, pick nominative or absolutive (if these are at all marked overtly). If neither the A nor the P argument of transitive clauses is identified as such by overt marking, or if case-marking is restricted to pronouns, assume the language has no “case”.

Within-language variation

- The family-bias methods allow a new approach to dealing with within-language variation:
- think of the structures within a language as logically independent of each other, forming the ultimate leaves of a family tree ("ultra-radical construction grammar"):



- and analyze family biases for the individual structures in exactly the sam we as we did for per-language data!

Models and theories

- Distributional biases can be modeled as generalized linear models, allowing tests of which factors best explain the data
- Statistical models ≠ causal models!
- Instead, statistical models need theoretical interpretation and motivation, i.e. typological theories
- In return, a typological theory is testable iff we can derive from it a set of statistical models, with well-defined variables p and q.
- Testable typological theories explain what's where.

Two main classes of typological theories (as I see it)

1. ‘Match’ (naturalist, functionalist) theories: some distributions are more likely than others because they are more ‘natural’, i.e. better tuned to the way our brain processes language and to our cognitive abilities
    - typical variables in naturalistic models: structural and discourse properties, perhaps also types of social structures (or cognitive models of them)
2. ‘Spread’ (replicationalist) theories: distributions reflect the spread of structures in time and space, i.e. descent and language contact
    - typical variables in resulting models: structural and discourse properties, socio-geographic areas; family skewing independent of predictors

- The most interesting research designs combine variables from both theory classes in one model!

Match theories

- Key idea: some distributions match better ...
  - the nature of the brain
  - the nature of communication
  - the nature of society
- A classical example: according to Hawkins, the nature of incremental processing is better matched by OV structures with A≠O coding than by OV structures with A=P in morphology. We can *derive from this* the testable model:
- $\log(\text{Freq}) = \alpha + \beta\text{CASE} + \beta\text{ORDER} + \beta\text{CASE}\cdot\text{ORDER}$
  - for which we found robust statistical support.

Match theories and relativism

- *A common misunderstanding:* Match theories must posit universal ‘natures’, i.e. ‘communication’ and ‘society’ must be in the singular.
- *But:* a distribution can universally match universal cognition just as well as it can variably match variable cognition. In either case, what is truly universal is the ‘match’ between a given type of cognition and a given linguistic distribution; the key point is that they **co-vary**: $q \sim p$!
- In fact and ironically, some of the best universal ‘matches’ come from research on linguistic relativity, where both cognition types and linguistic structures vary.

Overall conclusions

- Understanding typological distributions requires
  - developing explicit theories about explanatory factors:
    - ‘match’ factors: principles leading to universally uniform structural pressure in diachrony
    - ‘spread’ factors: effects of language contact and faithful inheritance within families
  - deriving statistical models from this and
  - testing these models against fine-grained variables in sufficiently rich databases,
  - using the same tool set as any other discipline,
  - and not artificially reducing diversity beforehand

Some common misconceptions

- “Typology is about synchronic classification of languages.”
  - ✶ *No, it is about measuring similarities of linguistic structures across and within languages and explaining these similarities in terms of diachronic factors*
- “Typology is functionalist.”
  - ✶ *Only true for ‘match’ theories; not for ‘spread’ theories, and both are important!*
- “Typology doesn’t care about formal (mathematical) models.”
  - ✶ *No, at least Multivariate Typology relies on the mathematical modeling of similarities and of distributions.*

Some common misconceptions

- “Typology doesn’t care about within-language diversity.”
  - ✴ *No!*
- “Typologists exaggerate diversity because they taken everything at face value, with no deep analysis!”
  - ✴ *No, but diversity is a quantitative and probabilistic issue, not a yes-or-no question*
- “The object of inquiry in linguistics is the human ability to acquire and use a natural language”
  - ✴ *No, the object of inquiry in linguistics is the distribution of linguistic structures; human abilities are the object of comparative psychology (and much more successfully so!)*