Исследование вариативности в Зиловском андийском

Г. А. Мороз (с участием С. Ферхеес)
Международная лаборатория языковой конвергенции, НИУ ВШЭ

05–06 апреля 2022, НИУ ВШЭ, Нижний Новгород
Корпусные технологии и компьютерные науки в гуманитарном знании (КонКорт-2022)

Plan

Исследование вариативности в зиловском диалекте андийского языка

Зиловские данные

Исследование нахско-дагестанских исследователей

Что если 10^5 “среднестатистических” исследователя ... приедет в Зило?

Заключение

Вариативность

- “Two equally interesting questions are at the heart of this book: how an extraordinary degree of idiosyncratic linguistic variation can coexist with an extraordinarily homogeneous speaker population, and how linguists might overlook the possibility of their coexistence.” [Dorian 2010: 3]

Вариативность

- “Two equally interesting questions are at the heart of this book: how an extraordinary degree of idiosyncratic linguistic variation can coexist with an extraordinarily homogeneous speaker population, and how linguists might overlook the possibility of their coexistence.” [Dorian 2010: 3]

- Я сейчас представлю результаты анализа вариативность в моноэтничном селении Зило (андийский язык), а также покажу, как мы пробовали оценить, как “среднестатистический” исследователь получил бы похожие результаты.

Данные

Данные были собраны у:

- 44 носителей андийского языка (нахско-дагестанская семья) во время полевого исследования (Ботлихский район, Дагестан) в 2019 году

Данные

Данные были собраны у:

- 44 носителей андийского языка (нахско-дагестанская семья) во время полевого исследования (Ботлихский район, Дагестан) в 2019 году

- и 23 исследователей нахско-дагестанских языков при помощи онлайн опроса.

Plan

Исследование вариативности в зиловском диалекте андийского языка

Зиловские данные

Исследование нахско-дагестанских исследователей

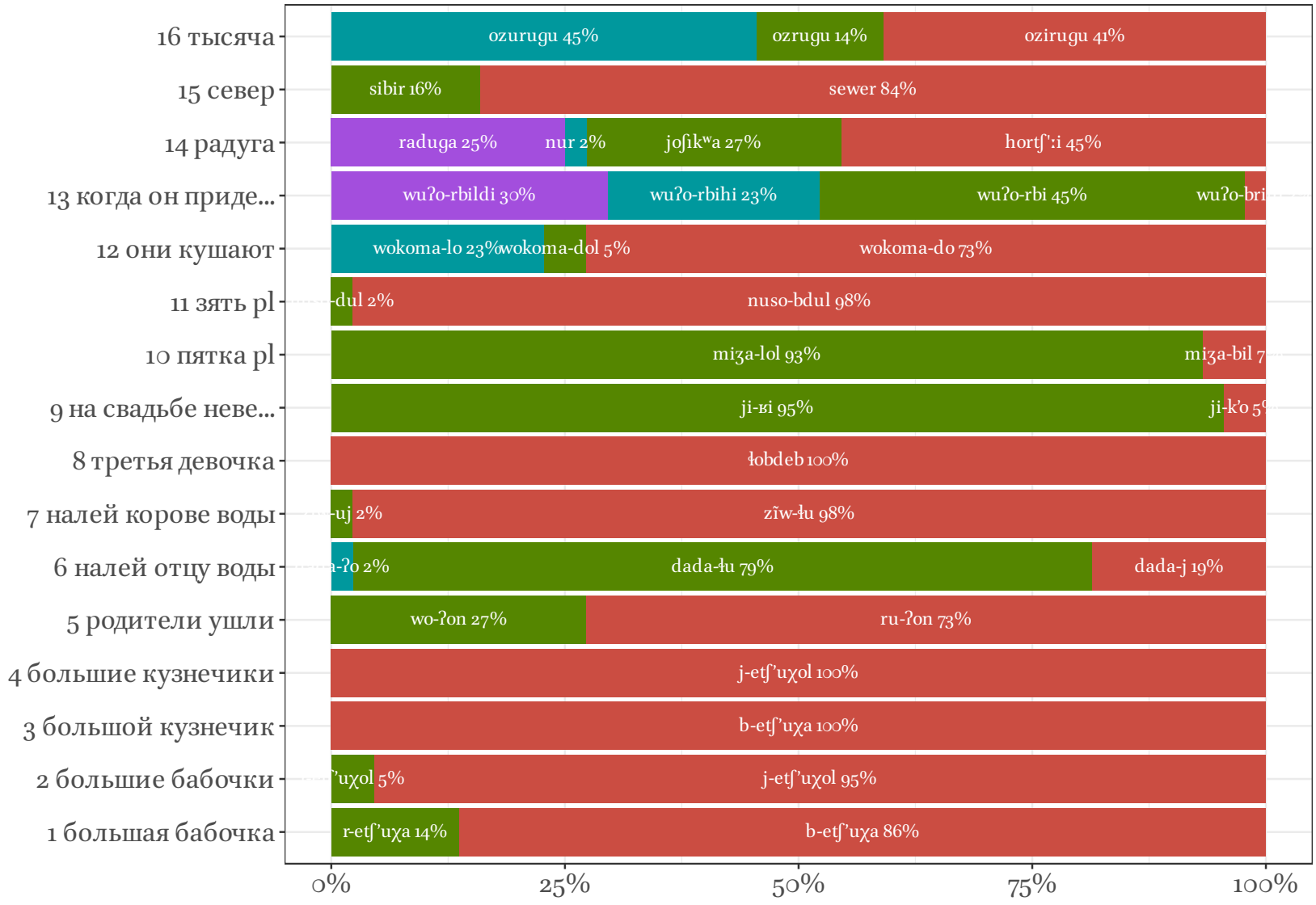
Что если 10^5 “среднестатистических” исследователя ... приедет в Зило?

Заключение

44 носителей зиловского перевели следующие предложения:

1. ‘большая бабочка’
2. ‘большие бабочки’
3. ‘большой кузнечик’
4. ‘большие кузнечики’
5. ‘родители ушли’
6. ‘налей отцу воды’
7. ‘налей своей корове воды’
8. ‘третья девочка’
9. ‘на свадьбе невеста была красивая’
10. ‘пятки’
11. ‘зятья’
12. ‘они едят’
13. ‘когда он придет, мы будем есть’
14. ‘радуга’
15. ‘север’
16. ‘тысяча’

Результаты зиловского опроса (44 носителей)



Информационная энтропия

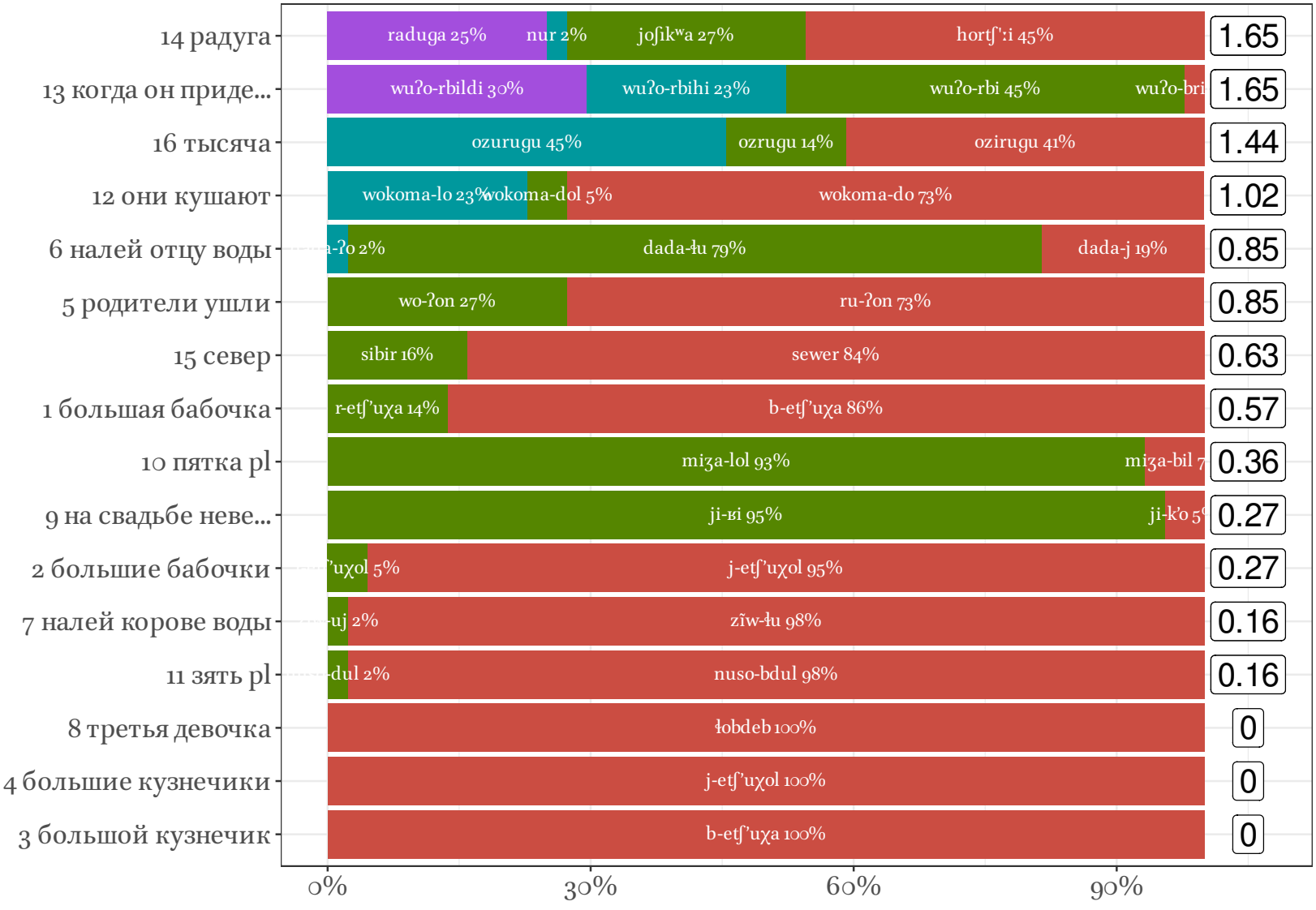
Чтобы измерить вариативность каждого вопросы, мы решили использовать информационную энтропию, введенную в [Shannon 1948]:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \times \log_2 P(x_i)$$

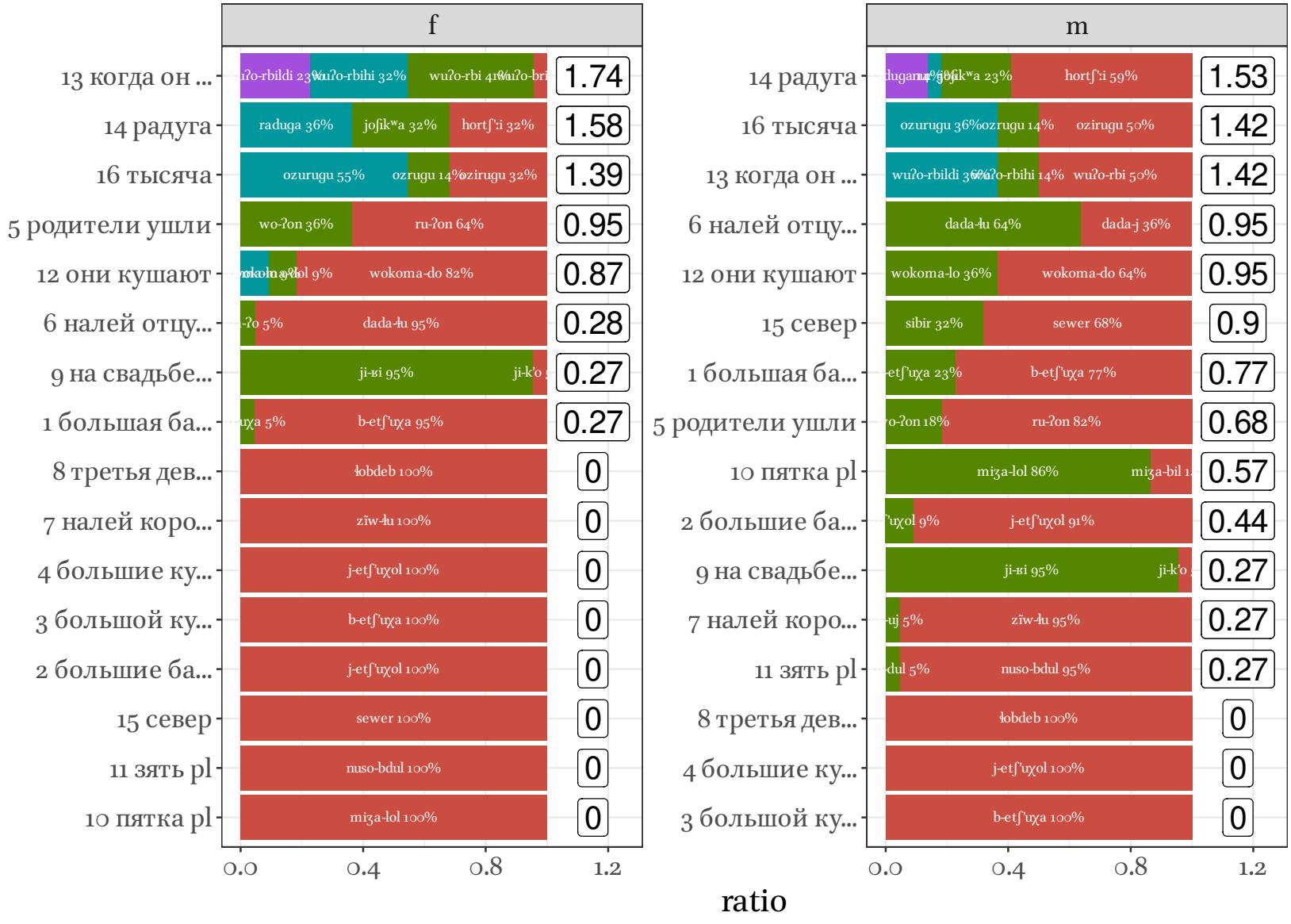
Область значения энтропии $H(X) \in [0, +\infty]$:

данные	энтропия
A-A-A-A-A	0.00
A-A-A-A-B	0.72
A-A-A-B-B	0.97
A-A-B-B-B	0.97
A-A-B-B-C	1.52
A-B-C-A-B	1.52

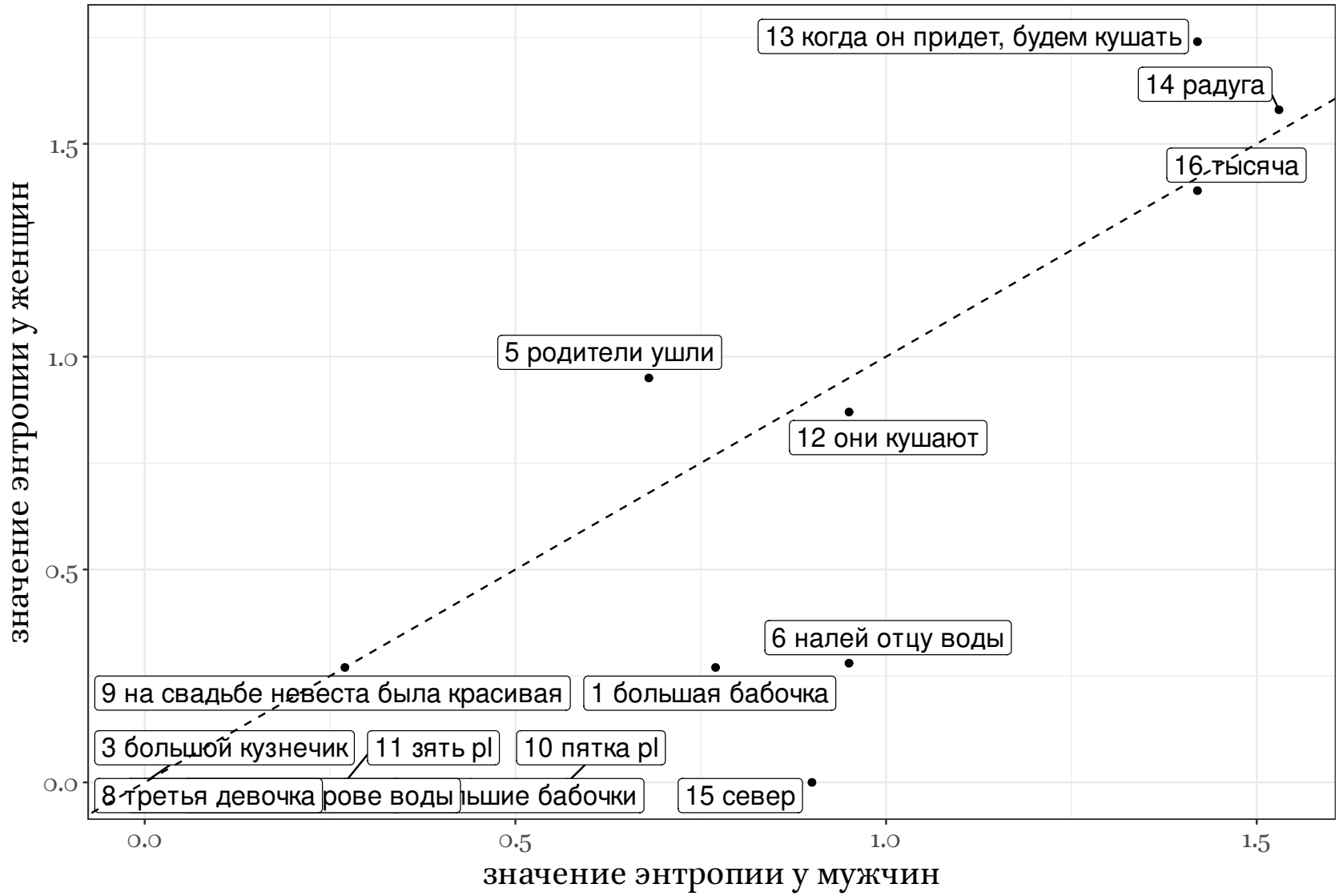
Зиловский опрос (44 носителей): значение энтропии справа



Зиловский опрос (44 носителей): гендерные различия



Зиловский опрос (44 носителей): значения энтропии в зависиимости от гендера



пунктирная линия -- y = x

Plan

Исследование вариативности в зиловском диалекте андийского языка

Зиловские данные

Исследование нахско-дагестанских исследователей

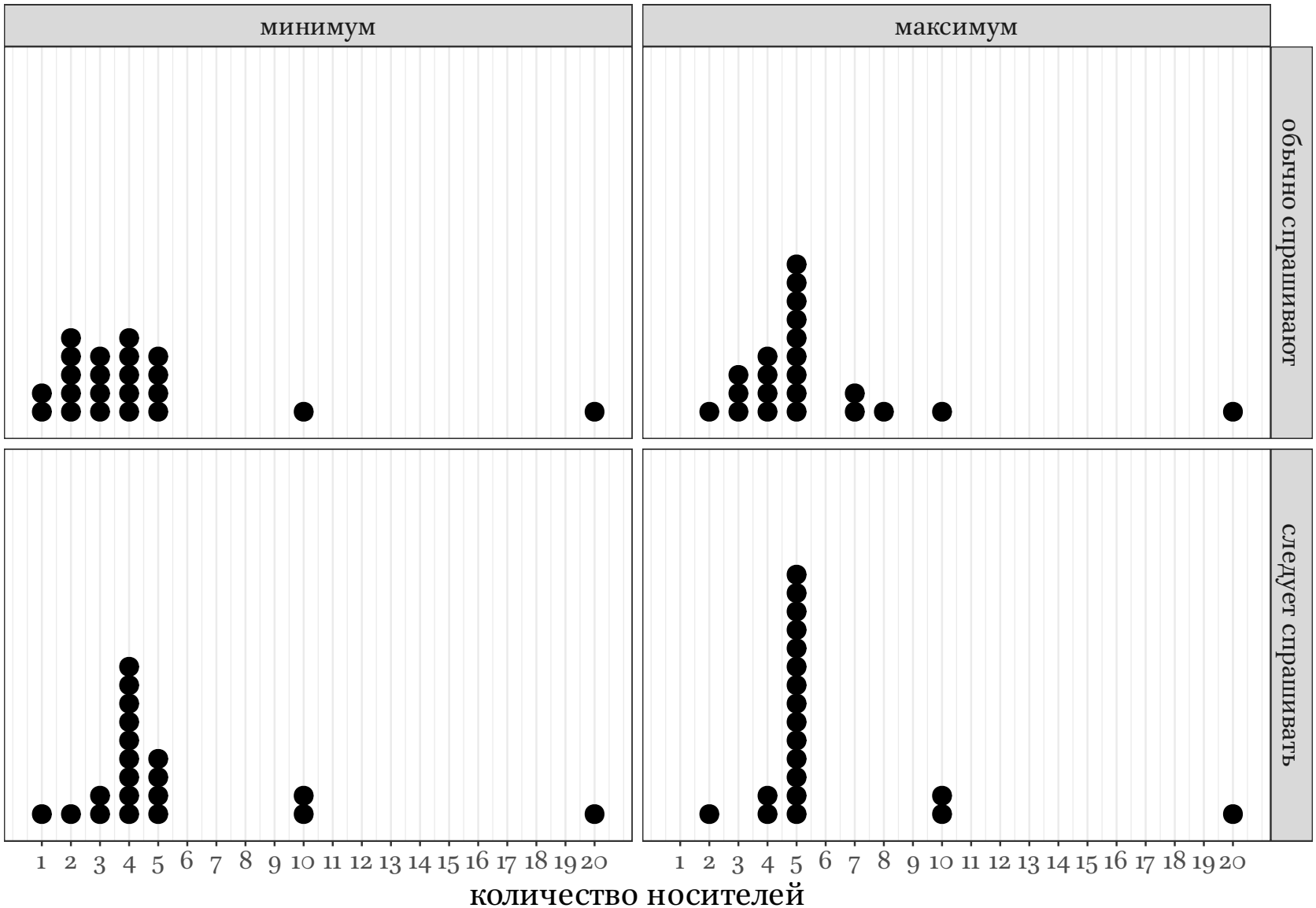
Что если 10^5 “среднестатистических” исследователя ... приедет в Зило?

Заключение

23 нахско-дагестанских исследователей заполнили следующую анкету:

- образование
- лингвистические интересы
- изучалась ли лингвистика в университете
- участие в полевой работе в качестве студента
- год получения степени
- место учебы/работы
- предпочтительное количество людей в полевой работе
- цели полевой работы
- **количество носителей, которые, согласно мнению исследователя, *следует* опрашивать**
- **количество носителей, которые исследователь *обычно* опрашивает**
- …

Количество носителей

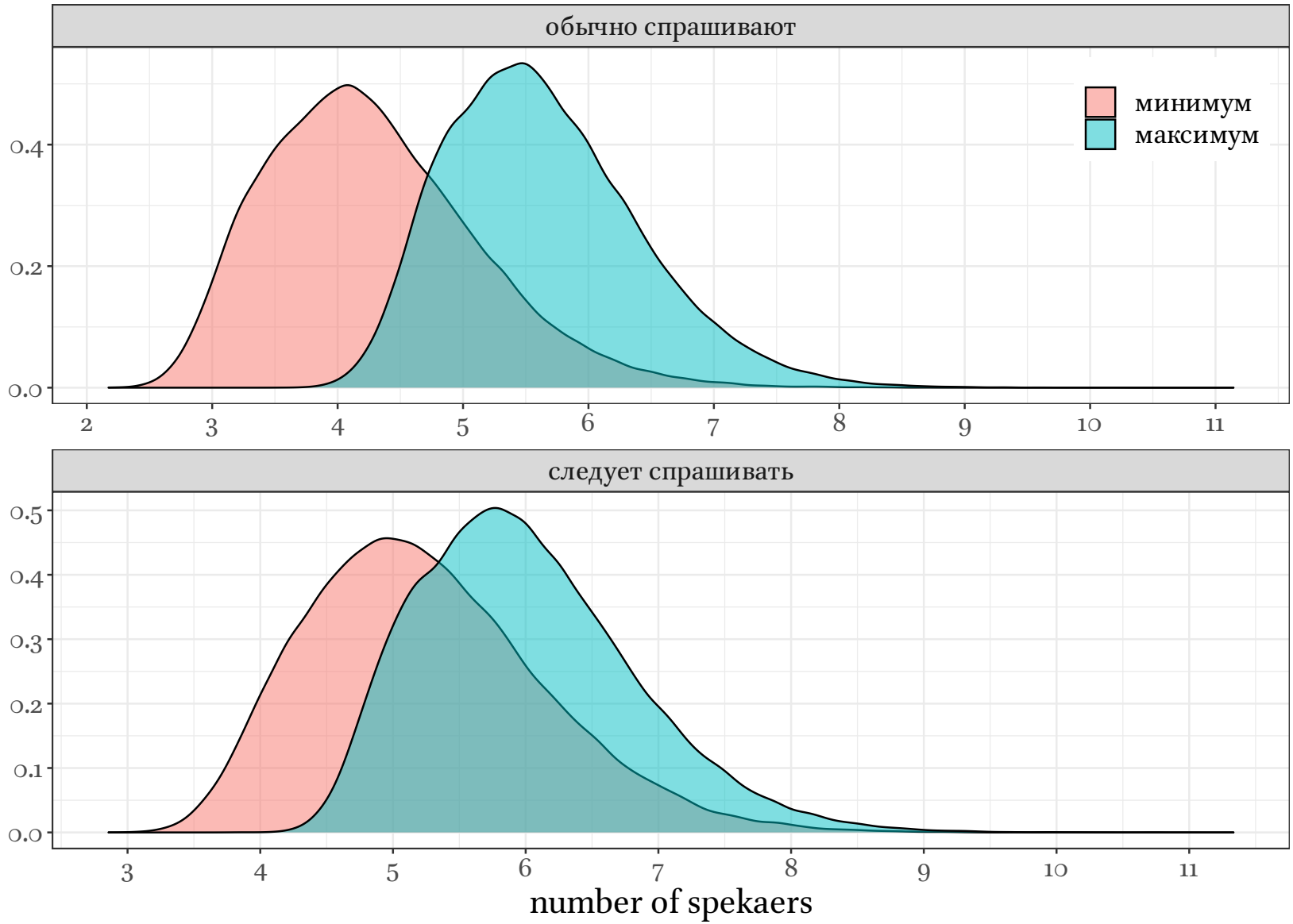


Бутстрэп

“To pull oneself over a fence by one’s bootstraps”.

Бутстрэп – это такой статистический подход, в рамках которого некоторый статистический параметр оценивается на основе большого количества выборок из имеющихся данных с повторением (т. е. каждое наблюдение может встретиться в выборке 0 раз, 1 раз, 2 раза и т. д.). В результате, вместо одной оценки параметра получается столько оценок, сколько у нас выборок, а все эти оценки формируют распределение.

Бустрэп среднего количества опрашиваемых носителей (10^5 iterations)



Plan

Исследование вариативности в зиловском диалекте андийского языка

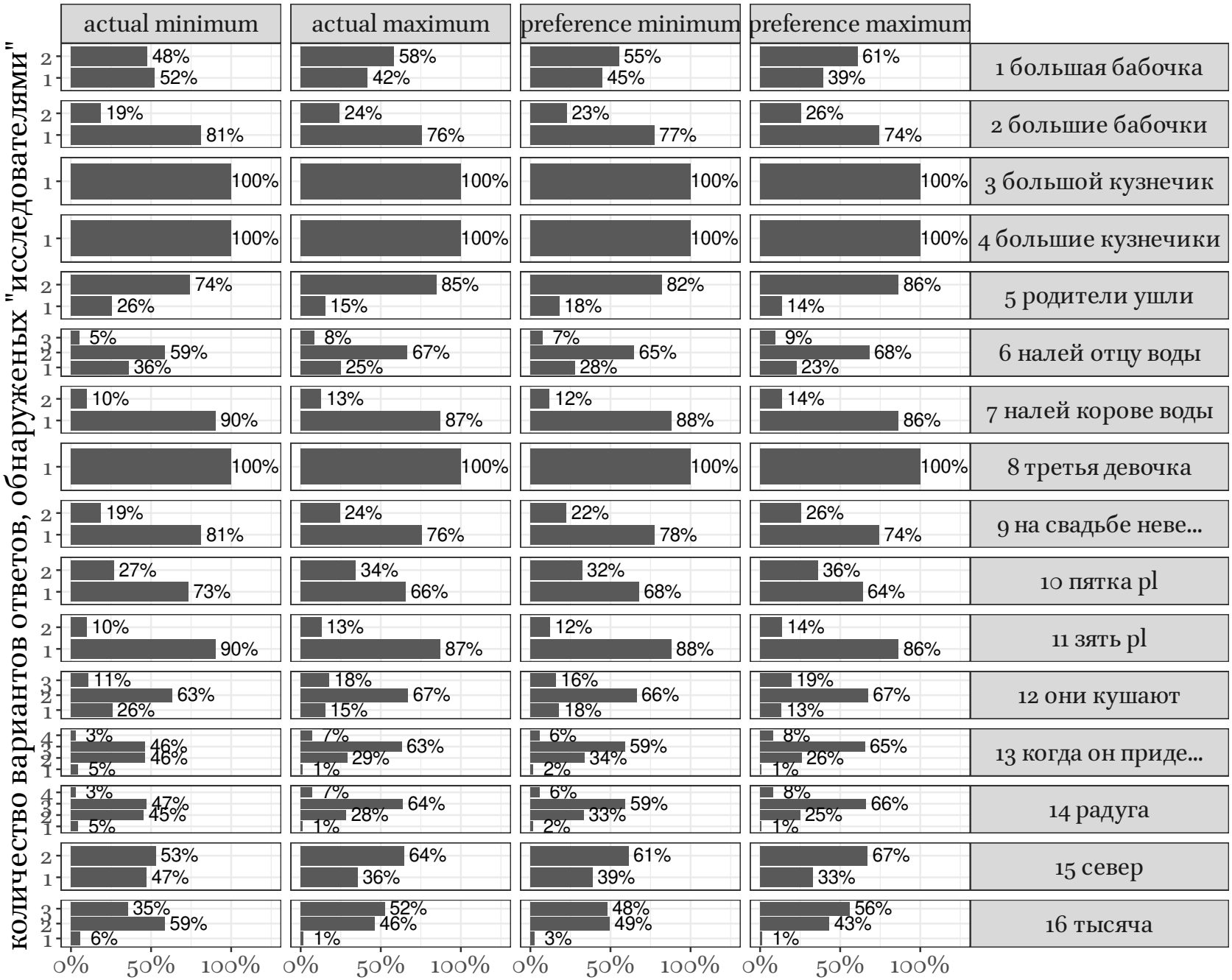
Зиловские данные

Исследование нахско-дагестанских исследователей

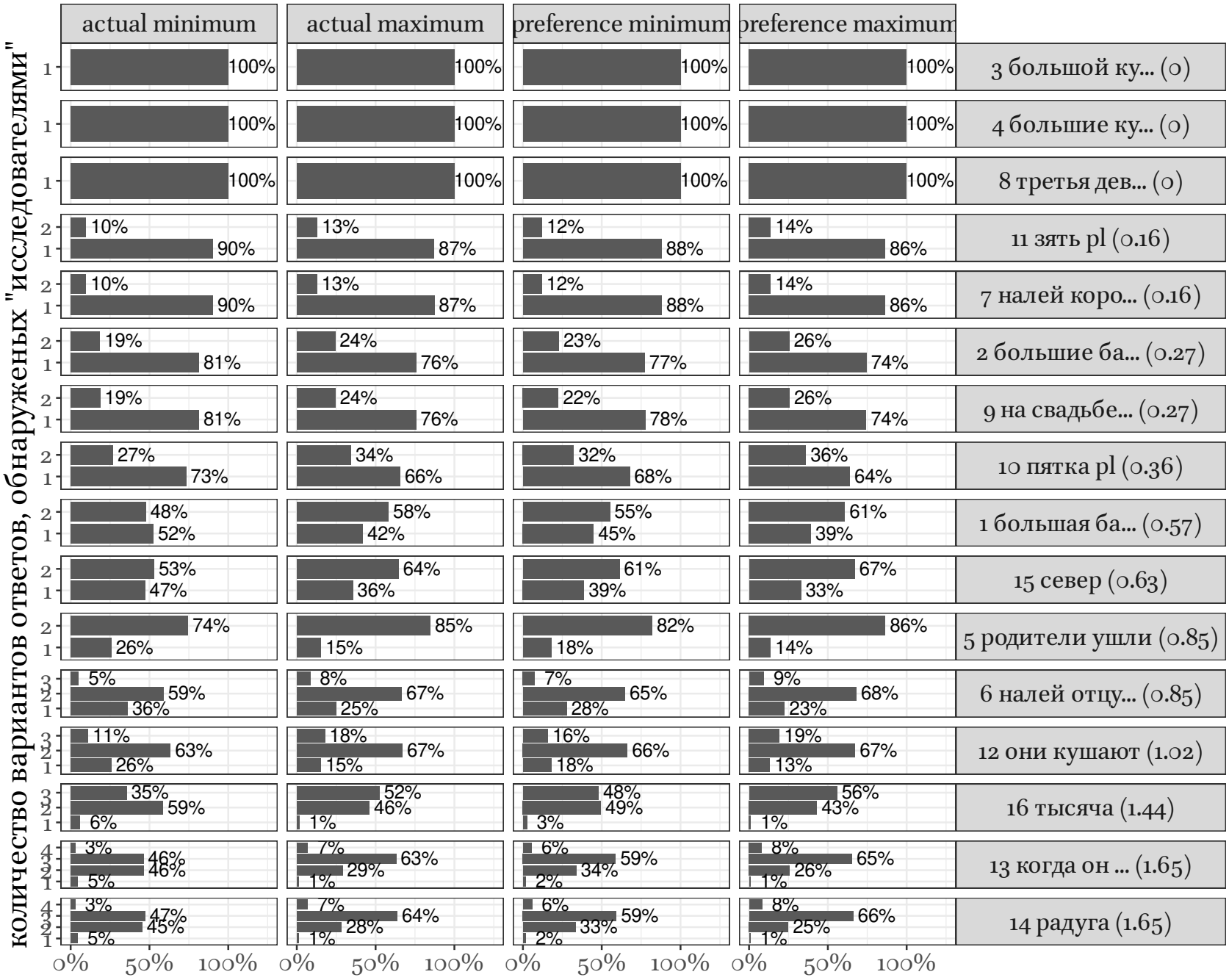
Что если 10^5 “среднестатистических” исследователя ... приедет в Зило?

Заключение

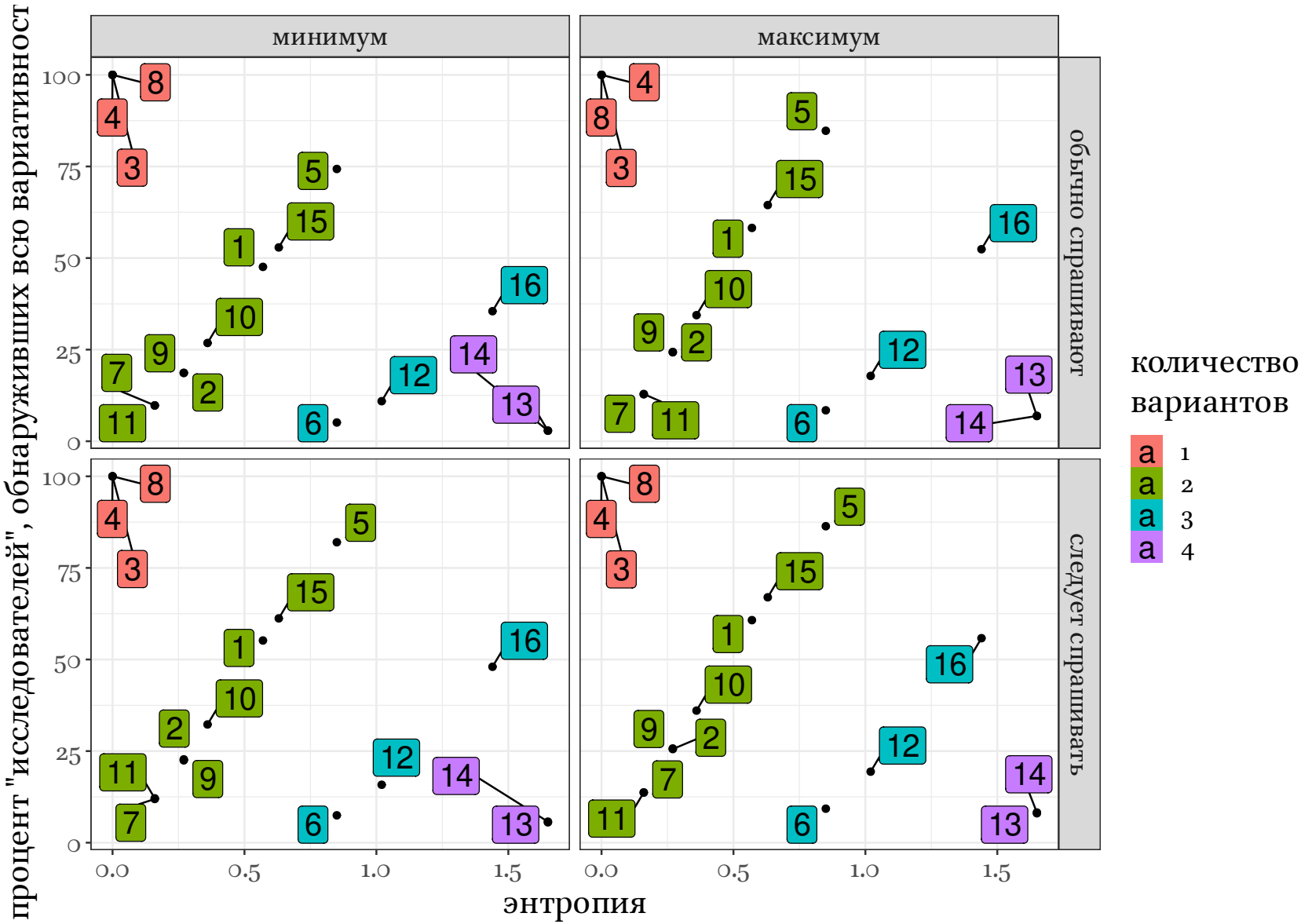
10^5 выборок из данных эксперимента



10^5 выборок из данных эксперимента (энтропия)



Когда “исследователи” найдут меньше?



Номер на графике соотносится с номером вопроса в анкете.

Plan

Исследование вариативности в зиловском диалекте андийского языка

Зиловские данные

Исследование нахско-дагестанских исследователей

Что если 10^5 “среднестатистических” исследователя ... приедет в Зило?

Заключение

Заключение

- вариативность можно описывать при помощи энтропии
- “среднестатистического” исследователя — осмысленная единица метаанализа, которую следует дальше исследовать
- естественно: количество обнаруженной любыми исследователями зависит от энтропии вопроса

References

Dorian, N. C. (2010). *Investigating variation: The effects of social organization and social setting*. Oxford University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.