

# КОНТРАСТИВНЫЙ АНАЛИЗ

## КВАНТИТАТИВНЫЙ АНАЛИЗ ТЕКСТА

---

Кирилл Александрович Маслинский

21.02.2022 / 03

ЕУ СПб

## КЛЮЧЕВЫЕ СЛОВА

---

## Ключевые слова

Использование контрастного корпуса

Отношение правдоподобия

Логарифм отношения шансов

Взвешенное отношение шансов

Задача — извлечение лексики, характерной для данного корпуса

- Контрастный корпус (reference corpus) — отражает словоупотребление в языке вообще или в более широкой предметной области
- Составить частотные списки слов для изучаемого и контрастного корпуса
- Отсортировать слова по расхождению частотности с ожидаемой на основании контрастного корпуса
- Ключевые слова изучаемого корпуса — наверху списка

### Simple maths (by Adam Kilgarriff)

«это слово встречается в этом корпусе вдвое чаще, чем в том»

- Самый простой подход
  - Нормализовать частотности
    - употреблений на тысячу или употреблений на миллион (IPM)
  - Вычислить отношение нормализованных частотностей
  - Отсортировать список слов по значению отношения

Для примера:

- Два корпуса по миллиону токенов
- Нормализовать частотности не нужно

**fc** focus corpus — изучаемый корпус

**rc** reference corpus — контрастный корпус

## ПРОБЛЕМА 1: НЕЛЬЗЯ ДЕЛИТЬ НА 0

слово	fc	rc	отношение
редкость	10	0	?
помешивать	100	0	?
вкуснотища	1000	0	?

Стандартное решение: прибавить 1:

слово	fc	rc	отношение
редкость	11	1	11
помешивать	101	1	101
вкуснотища	1001	1	1001

## ПРОБЛЕМА 2: ИЗ-ЗА РЕДКИХ СЛОВ СЛИШКОМ МНОГО БОЛЬШИХ ОТНОШЕНИЙ

Частотность тоже важна. Решение: прибавить  $n$ .

•  $n = 1$

слово	fc	rc	fc+n	rc+n	отношение
изредка	10	0	11	1	11,00
временами	200	100	201	101	1,99
часто	12000	10000	12001	10001	1,20

•  $n = 100$

слово	fc	rc	fc+n	rc+n	отношение
временами	200	100	300	200	1,50
изредка	10	0	110	100	1,10
часто	12000	10000	12100	10100	1,20



## Ключевые слова

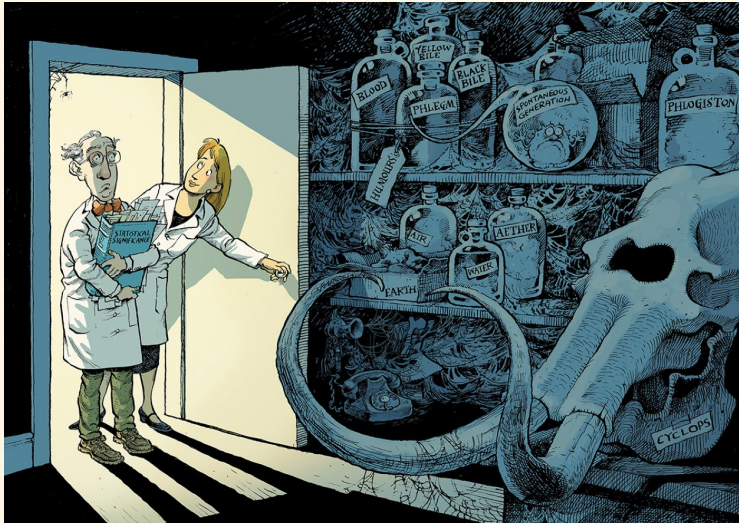
Использование контрастного корпуса

Отношение правдоподобия

Логарифм отношения шансов

Взвешенное отношение шансов

## УХОДЯЩАЯ ЭПОХА СТАТИСТИЧЕСКОЙ ЗНАЧИМОСТИ

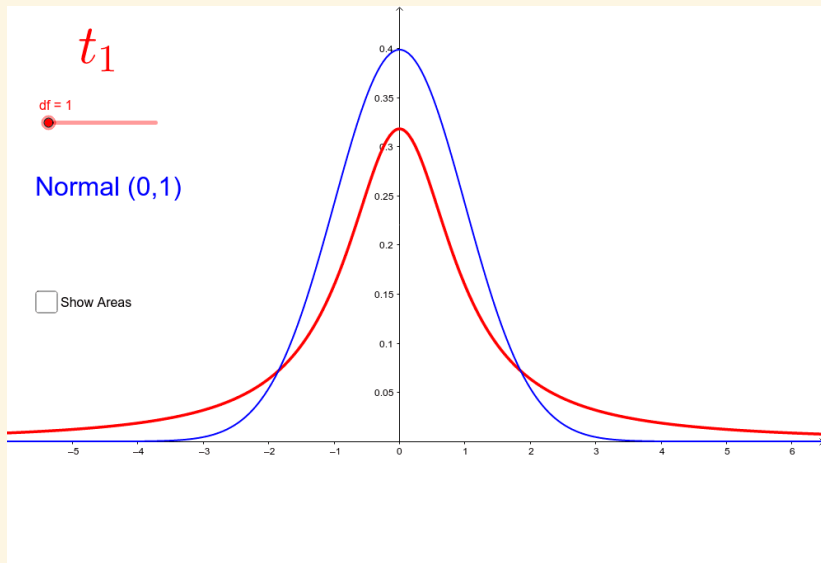


Amrhein et al. "Scientists rise up against statistical significance" (2019)

В парадигме стандартных статистических тестов при сравнении частотностей слов возникали проблемы:

- Предположение о нормальности неверно в случае частотного распределения слов
- В языке слишком много редких событий
- Неприменимость тестов, основанных на предположении о нормальности (напр., хи-квadrat), как минимум к редким событиям (частотность  $< 5$ )

# НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ ИЗЛИШНЕ УДИВЛЕНО



Способ включить частотности слов в парадигму статистических тестов:

Ted Dunning “Accurate Methods for the Statistics of Surprise and Coincidence” (1994)

- Отношение правдоподобия менее зависит от предположения о нормальности распределения данных
- Поэтому не так резко завышает значимость редких событий и может применяться для оценки различий не только самых частотных слов

## ОТНОШЕНИЕ ПРАВДОПОДОБИЯ: ФОРМУЛА

	Корпус 1	Корпус 2	Всего
Частотность слова	a	b	a+b
Частотность остальных слов	c-a	d-b	c+d-a-b
Всего	c	d	c+d

Ожидаемые частотности:

$$E1 \frac{c}{c+d}(a+b)$$

$$E2 \frac{d}{c+d}(a+b)$$

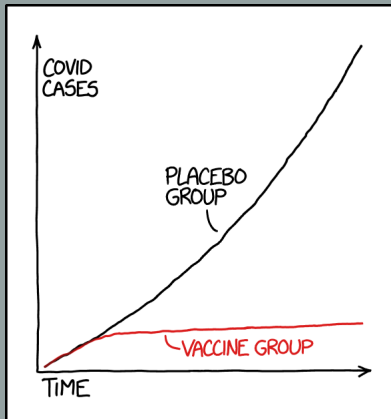
$$LL = G^2 = 2(a \log(a/E1) + b \log(b/E2)) \quad (1)$$

- Более чувствителен к частотным событиям (словам), чем к менее частотным [занижает степень различия по менее частотным словам]

# ЭКСПЕРИМЕНТ НА ЦАРЯХ И ЦВЕТОЧКАХ



## ПРАКТИЧЕСКИЙ ВЫВОД



STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

## Ключевые слова

Использование контрастного корпуса

Отношение правдоподобия

Логарифм отношения шансов

Взвешенное отношение шансов

шансы 1 : 3

вероятность  $1/4$  ,  $3/4$

шансы красных  $1 : 3 = 0,33$

вероятность красных  $1/4 = 0,25$

шансы через вероятность  $p/(1 - p) =$   
 $0,25/(1 - 0,25) = 1/4 : 3/4 = 0,33$

шансы 1 : 3

вероятность 1/4 , 3/4

шансы красных 1 : 3 = 0,33

вероятность красных 1/4 = 0,25

шансы через вероятность  $p/(1 - p) =$   
 $0,25/(1 - 0,25) = 1/4 : 3/4 = 0,33$

## ДВА СОБЫТИЯ: ОТНОШЕНИЕ ШАНСОВ

---

$$\text{шансы1 } 1 : 3 = 0,33$$

$$\text{шансы2 } 2 : 5 = 0,4$$

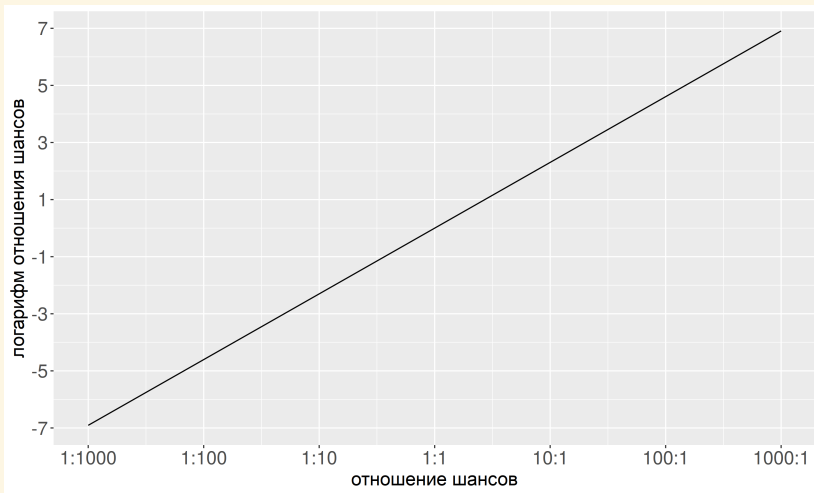
$$\text{шансы1/шансы2 } 1/3 : 2/5 = 5/6 = 0,83$$

$$\text{шансы2/шансы1 } 2/5 : 1/3 = 6/5 = 1,2$$

$$\text{шансы1/шансы2 } \log(5/6) = -0,18$$

$$\text{шансы2/шансы1 } \log(6/5) = 0,18$$

# ШАНСЫ НА ЛОГАРИФМИЧЕСКОЙ ШКАЛЕ



## Ключевые слова

Использование контрастного корпуса

Отношение правдоподобия

Логарифм отношения шансов

Взвешенное отношение шансов



$$\delta = \frac{\frac{f_{(w,c1)} + \alpha_{(w,c1)}}{N_{c1} + \alpha_{c1} - f_{(w,c1)} - \alpha_{(w,c1)}}}{\frac{f_{(w,c2)} + \alpha_{(w,c2)}}{N_{c2} + \alpha_{c2} - f_{(w,c2)} - \alpha_{(w,c2)}}}$$