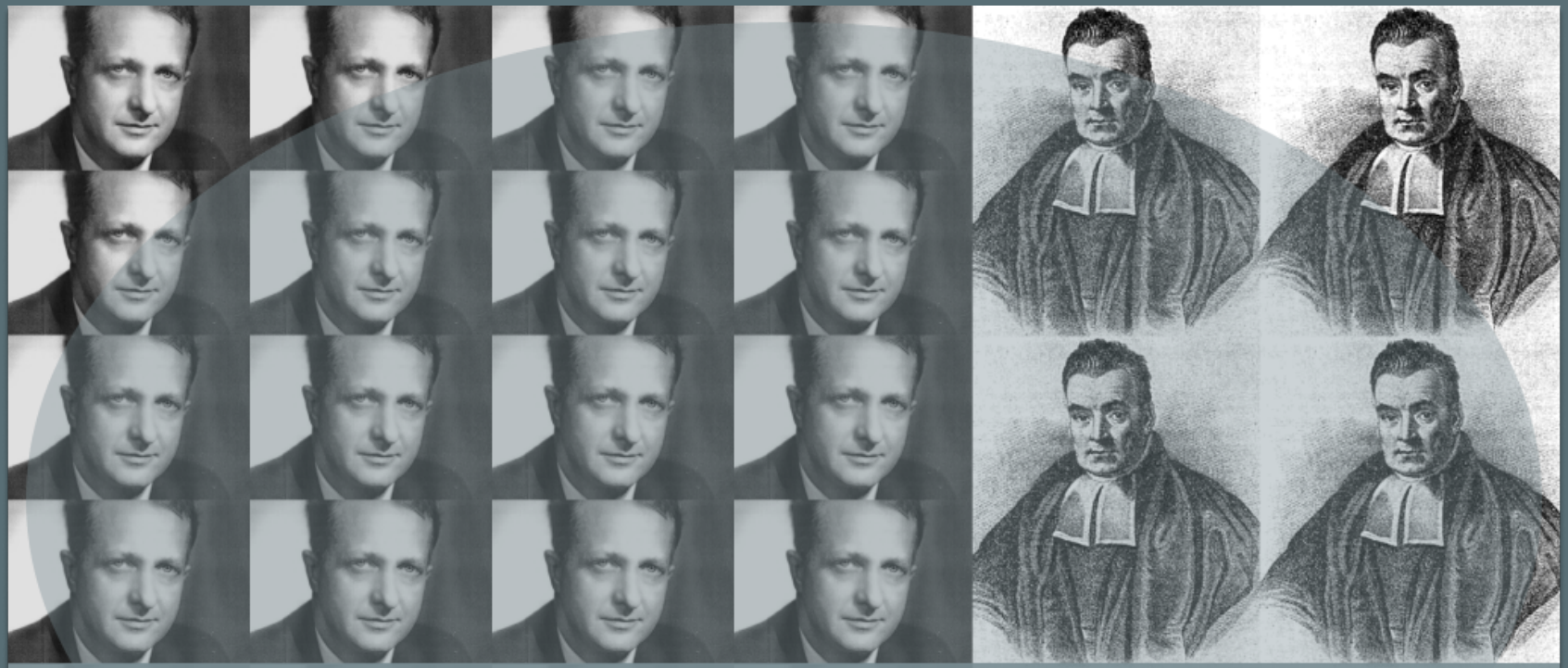


Кирилл Маслинский

# «Встреча Байеса и Ципфа, или Как измерить различия в частотности слов»

27 января 2024 года

# Задача

Предположим, у нас есть несколько групп текстов, которые различаются: жанром, временем или местом создания, полом авторов, строгостью цензуры или чем-то еще. Посчитав слова в каждой группе, мы непременно найдем отличия в частотности.

- Когда разброс в частотности указывает на устойчивые различия между авторами текстов и условиями их создания?
- Как выделить интересные различия?
- Как в этом может помочь закон Ципфа и байесовские статистические методы?

# Метод контрастного корпуса

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

# Постулат 1

Вероятность появления слова постоянна в рамках корпуса

следовательно

Различия в частотности слова между корпусами мы можем трактовать как различия в вероятности слова в разных корпусах.

# Персонаж: Джордж Ципф

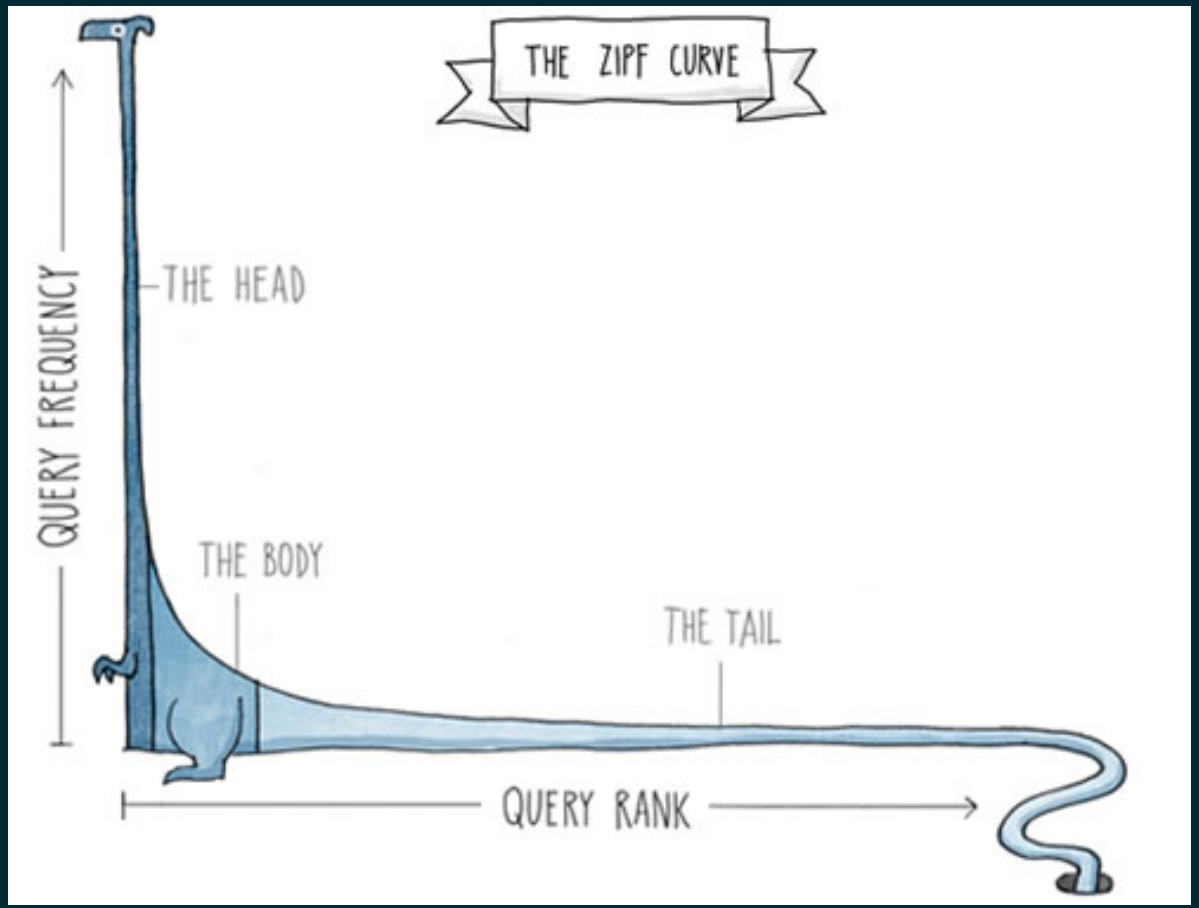
Закон Ципфа Предсказывает частотность слова по его рангу в частотном списке:

$$f(w) \sim \frac{1}{r(w)} \tag{1}$$

$f(w)$ — частотность слова $w$

$r(w)$ — ранг слова $w$ в частотном списке

# Персонаж: Джордж Ципф



# Эксперимент с царями и цветочками

# Отношение правдоподобия

# Dunning 1994

## Accurate Methods for the Statistics of Surprise and Coincidence

Ted Dunning*
New Mexico State University

*Much work has been done on the statistical analysis of text. In some cases reported in the literature, inappropriate statistical methods have been used, and statistical significance of results have not been addressed. In particular, asymptotic normality assumptions have often been used unjustifiably, leading to flawed results.*

*This assumption of normal distribution limits the ability to analyze rare events. Unfortunately rare events do make up a large fraction of real text.*

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

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

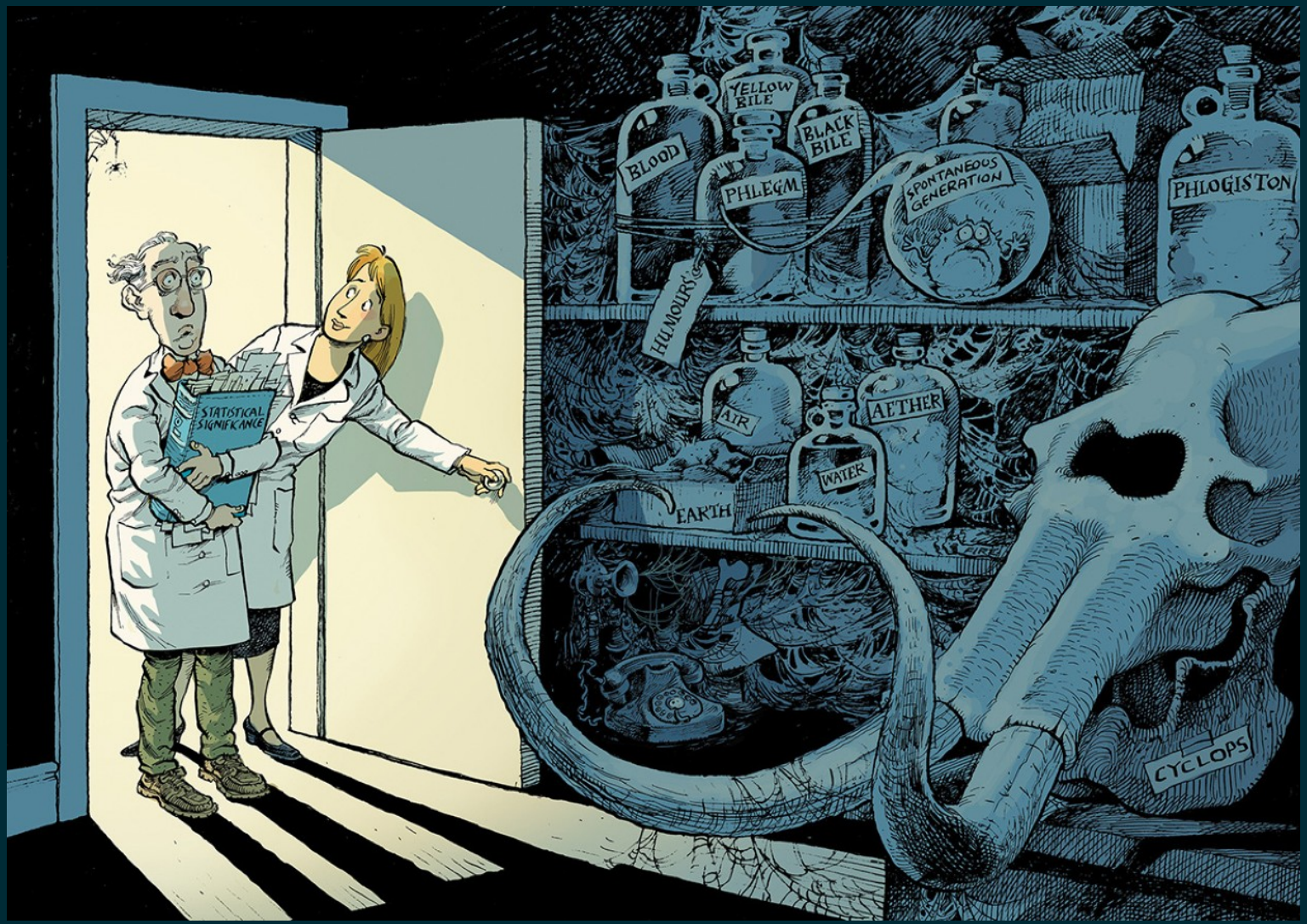
# Уровни значимости Log Likelihood

Цитата с сайта: `https://ucrel.lancs.ac.uk/llwizard.html`

The higher the G2 value, the more significant is the difference between two frequency scores. For these tables, a G2 of 3.8 or higher is significant at the level of $p < 0.05$ and a G2 of 6.6 or higher is significant at $p < 0.01$.

- 95th percentile; 5% level; $p < 0.05$; critical value = 3.84
- 99th percentile; 1% level; $p < 0.01$; critical value = 6.63
- 99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
- 99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13

# Уходящая эпоха статистической значимости




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

# Simple maths

# Простое отношение частотности

**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 | ранг |
|---|---|---|---|---|---|---|
| изредка | 10 | 0 | 11 | 1 | 11,00 | 1 |
| временами | 200 | 100 | 201 | 101 | 1,99 | 2 |
| часто | 12000 | 10000 | 12001 | 10001 | 1,20 | 3 |

- $n = 100$

| слово | fc | rc | fc+n | rc+n | отношение | ранг |
|---|---|---|---|---|---|---|
| временами | 200 | 100 | 300 | 200 | 1,50 | 1 |
| изредка | 10 | 0 | 110 | 100 | 1,10 | 3 |
| часто | 12000 | 10000 | 12100 | 10100 | 1,20 | 2 |

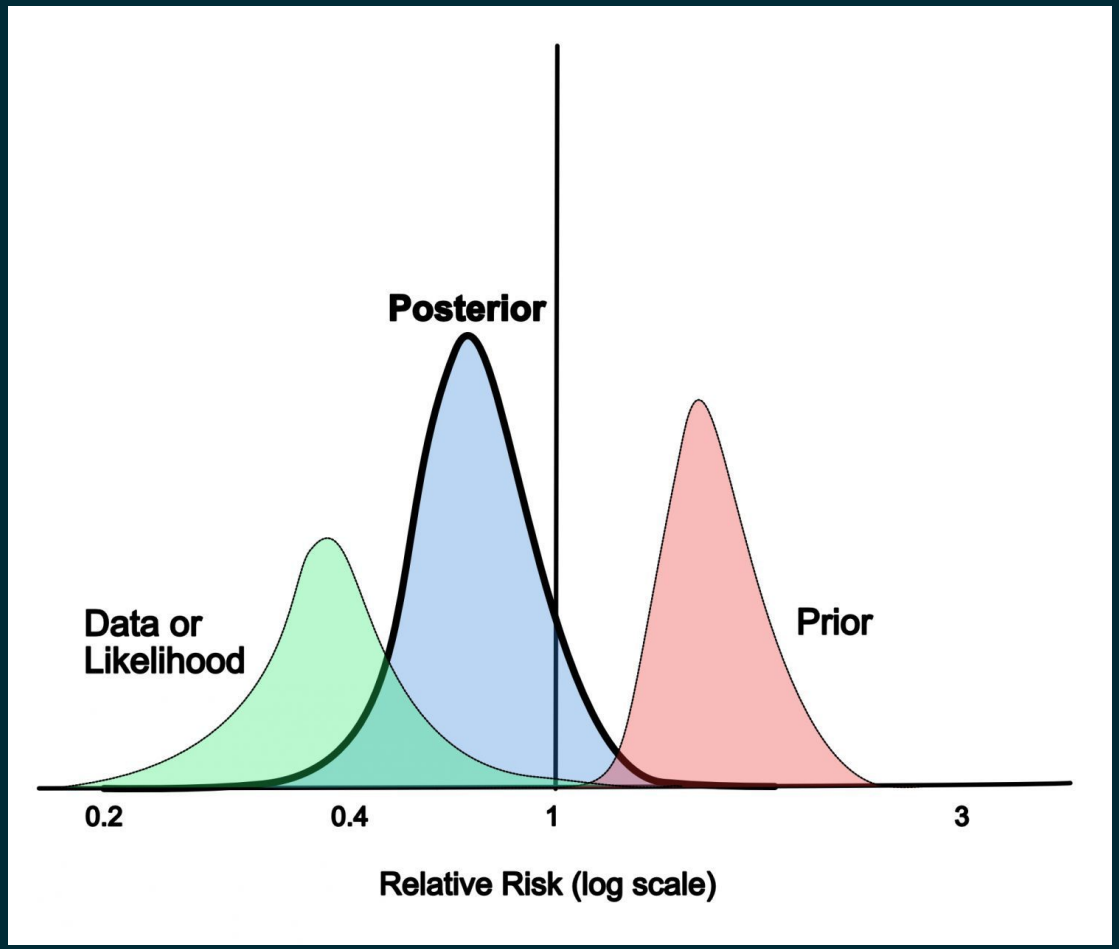
# Персонаж: Томас Байес

Теорема Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

$$posterior = \frac{likelihood \cdot prior}{evidence} \tag{3}$$

# Персонаж: Томас Байес



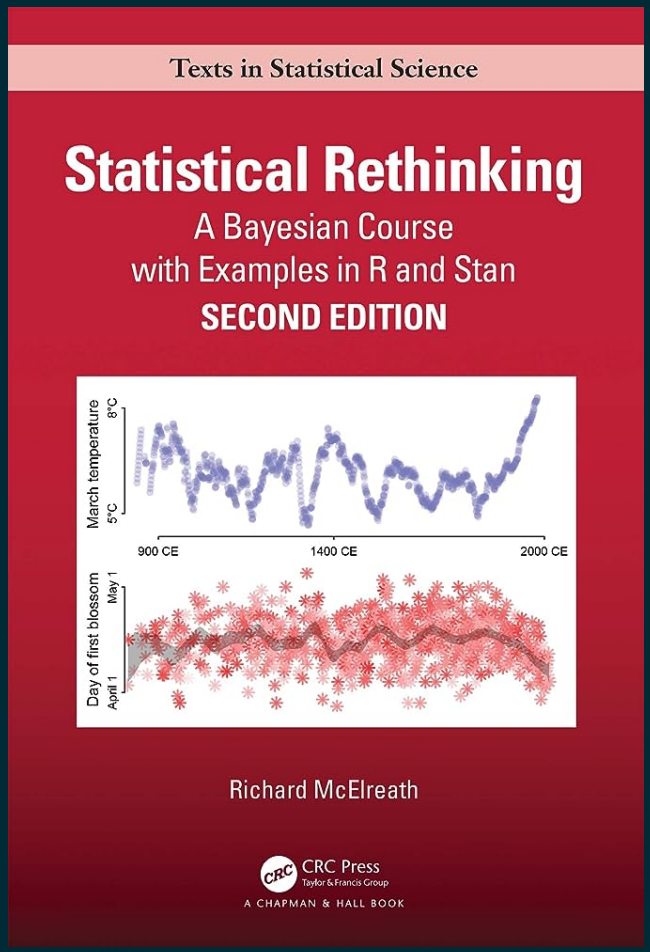

# Постулат 2

## Текстуальная гомогенность (пренебрегаем структурой текста)

следовательно

Можем рассматривать частотность слова в корпусе как результат *пуассоновского процесса* (с параметром $\lambda$, отражающим вероятность слова в корпусе).

# Statistical Rethinking by Richard McElreath




AND free lecture series on Youtube

# Take-aways

- Сравнение лексической частотности строится на имплицитной упрощенной порождающей модели
- Полезно делать постулаты этой модели эксплицитными
- И оценивать собственные методы на основании синтетических данных с известными параметрами
- Байес не делает чудес, но учит сознательности и смирению

# Take-aways

- Сравнение лексической частотности строится на имплицитной упрощенной порождающей модели
- Полезно делать постулаты этой модели эксплицитными
- И оценивать собственные методы на основании синтетических данных с известными параметрами
- Байес не делает чудес, но учит сознательности и смирению

# Данные и код примеров

https://github.com/maslinych/bayes-zipf