ТАНКИ, УТКИ И ШАРЫ

Сергей Николенко

СПбГУ — Санкт-Петербург

13 ноября 2026 г.





Random facts:

- 13 ноября 1833 г. над Северной Америкой наблюдали мощный метеорный шторм Леонид, ставший важной вехой в изучении метеоров
- 13 ноября 1851 г. по подводному телеграфному кабелю Дувр — Кале передали первое сообщение; кабель прослужил до 1875 года
- 13 ноября 1927 г. туннель Холланда под Гудзоном открылся для автомобильного движения между Нью-Йорком и Нью-Джерси
- 13 ноября 1940 г. в Нью-Йорке состоялась премьера «Фантазии» Диснея, соединившей анимацию и классическую музыку
- 13 ноября 1971 г. Mariner 9 вышел на орбиту Марса и стал первым искусственным спутником другой планеты
- 13 ноября 1982 г. в Вашингтоне торжественно открыли Мемориал ветеранов Вьетнама по проекту Майи Лин

Обучение равномерного распределения

Как подсчитать немецкие танки

- Как во время второй мировой понять, сколько танков произвела Германия в этом году?
- Есть традиционные средства разведки: оценить число немецких танков там, где мы смогли их увидеть, и как-то экстраполировать на всё остальное
- Но нашёлся и другой неожиданный путь: предположим, что вы захватили немецкий танк (например, вывезли подбитый с поля боя); оказалось, что на разных деталях танка дотошные немецкие производители… наносят серийные номера!
- Мы можем посмотреть на, условно говоря, коленвал очередной «Пантеры» и узнать, что это коленвал «Пантеры» номер 273
- Что это нам скажет?

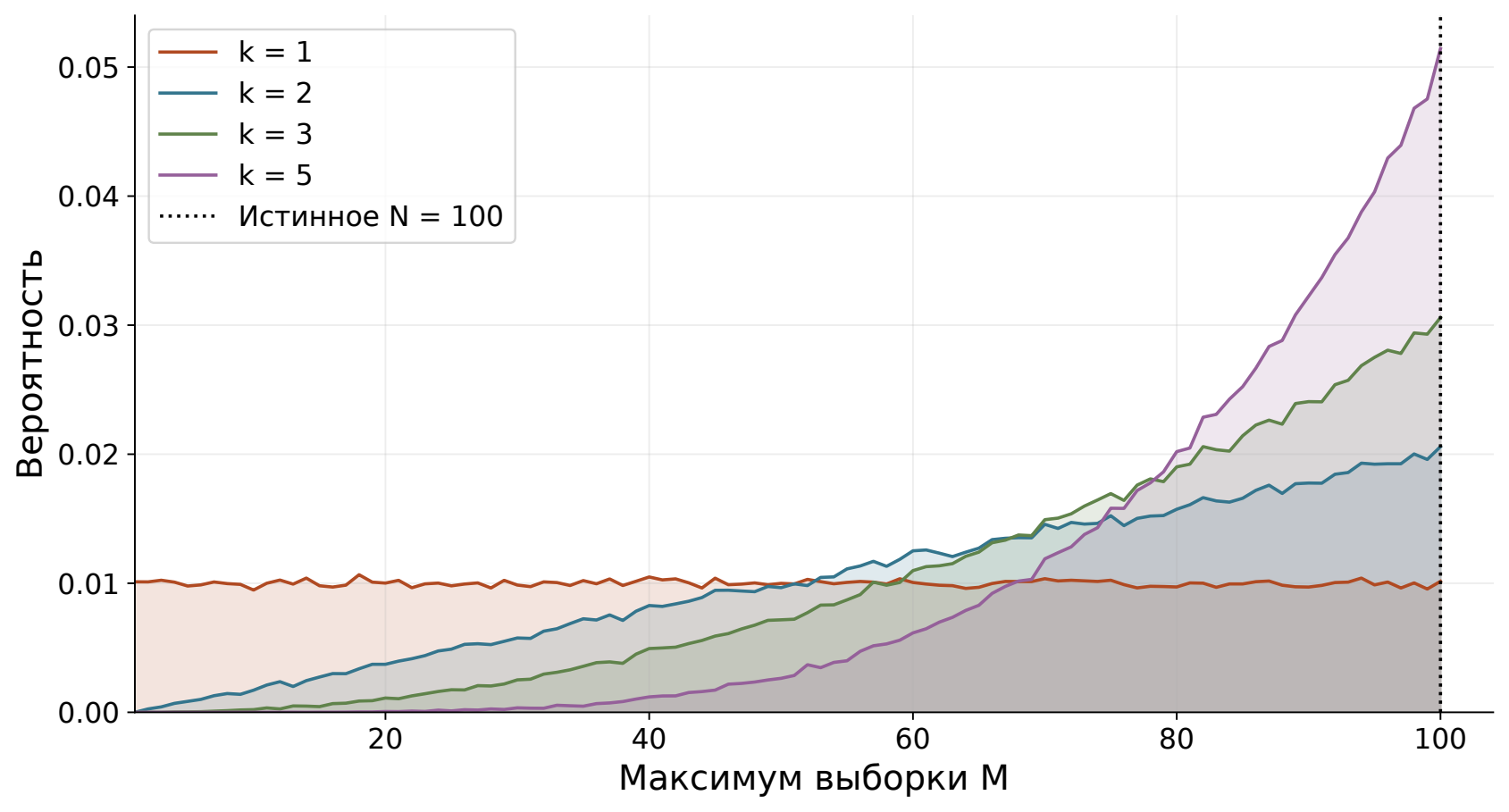
Как подсчитать немецкие танки

- Задача кажется несложной: даны несколько чисел, взятых из равномерного распределения на интервале $1, \dots, N$, и требуется оценить N
- Начнём с самого простого случая, когда число только одно; мы захватили танк с серийным номером m и хотим оценить, сколько всего их могло бы быть
- Функция правдоподобия:

$$p\left(m|N\right) = \begin{cases} \frac{1}{N}, & \text{если } N \geq m, \\ 0, & \text{если } N < m \end{cases}$$

- Гипотеза максимального правдоподобия — это $N = m$, тогда правдоподобие исхода m (и всех остальных) равно $\frac{1}{m}$
- Странный вывод, да? Это смещённая оценка, а нам бы несмещённую...

Распределение максимума выборки



Как подсчитать немецкие танки

- Ответ из статистики: найти распределение максимума M из выборки, то есть оценить $\mathbb{E}[M \mid N]$, а затем подставить в качестве этого ожидания эмпирический максимум $m = \max_{i=1}^{k} x_i$ и решить уравнение относительно N
- Для $k = 1$ (одно число в выборке) это совсем просто: ожидание максимума совпадает с ожиданием одного x_i и равно $\frac{N+1}{2}$, а значит,

$$\hat{N} = 2m - 1$$

- Это звучит вполне интуитивно: мы ожидаем, что равномерно взятое случайное число в среднем попадёт в середину интервала, значит, мы ожидаем, что сам интервал будет примерно вдвое шире этого числа

Как подсчитать немецкие танки

- Для $k > 1$ придётся заняться комбинаторикой; распределение $p\,(M|N)$ равно

$$p\,(M = m|N) = \frac{\text{число выборок из } k \text{ чисел с максимумом } m}{\text{общее число выборок из } k \text{ чисел}}$$

- В знаменателе число сочетаний из N по k, $\binom{N}{k}$, а в числителе, соответственно, число сочетаний из $m-1$ по $k-1$: мы уже зафиксировали одно из чисел m, и это максимум, то есть остальные $k-1$ число должны быть взяты от 1 до $m-1$
- Итого получается, что

$$p\,(M = m|N) = \frac{\binom{m-1}{k-1}}{\binom{N}{k}}.$$

Как подсчитать немецкие танки

- А ожидание максимума, значит, равно

$$\mathbb{E}\left[M \mid N\right] = \sum_{m=k}^{N} mp\left(M = m|N\right) = \sum_{m=k}^{N} m\frac{\binom{m-1}{k-1}}{\binom{N}{k}} = \frac{1}{\binom{N}{k}}\sum_{m=k}^{N} m\binom{m-1}{k-1} =$$

$$= \frac{1}{\binom{N}{k}}\sum_{m=k}^{N} m\frac{(m-1)!}{(k-1)!(m-k)!} = \frac{k}{\binom{N}{k}}\sum_{m=k}^{N}\frac{m!}{k!(m-k)!} = \frac{k}{\binom{N}{k}}\sum_{m=k}^{N}\binom{m}{k};$$

  в предпоследнем равенстве мы умножили и разделили на k
- Воспользуемся известным тождеством о биномиальных коэффициентах: $\sum_{a=c}^{b}\binom{a}{c} = \binom{b+1}{c+1}$; его легко доказать индукцией по b: для $b = c$ слева и справа единица, а если мы уже доказали тождество для некоторого b, то индукционный переход выглядит так:

$$\sum_{a=c}^{b+1}\binom{a}{c} = \sum_{a=c}^{b}\binom{a}{c} + \binom{b+1}{c} = \binom{b+1}{c+1} + \binom{b+1}{c} = \binom{b+2}{c+1};$$

  последнее равенство — свойство треугольника Паскаля

Как подсчитать немецкие танки

- Применяя доказанное тождество, получим, что

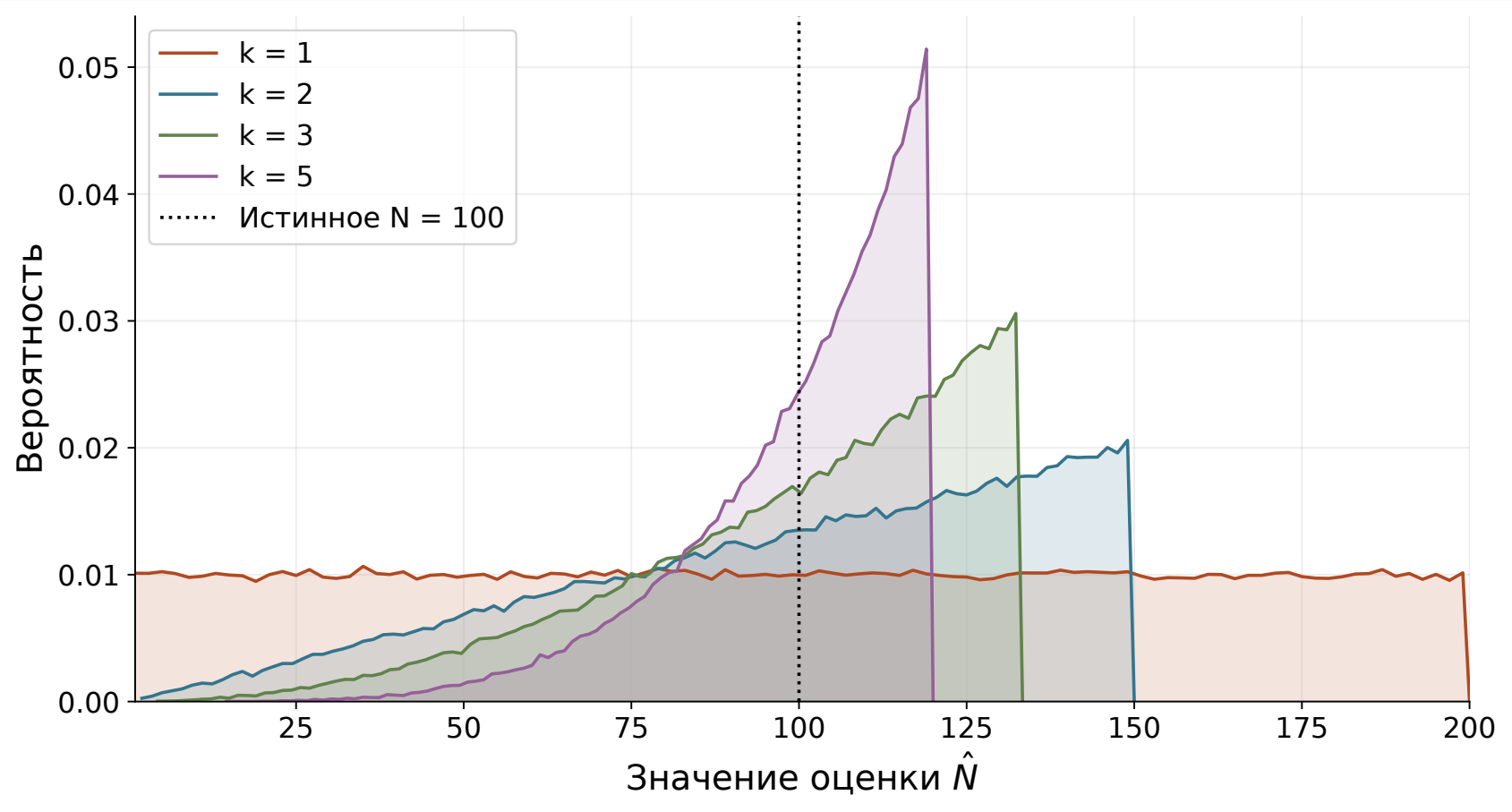
$$\mathbb{E}\left[M \mid N\right] = \frac{k}{\binom{N}{k}}\binom{N+1}{k+1} = \\ = k\frac{k!(N-k)!}{N!}\frac{(N+1)!}{(k+1)!(N-k)!} = (N+1)\frac{k}{k+1}.$$

- Получилась очень простая формула, которую легко теперь и обратить:

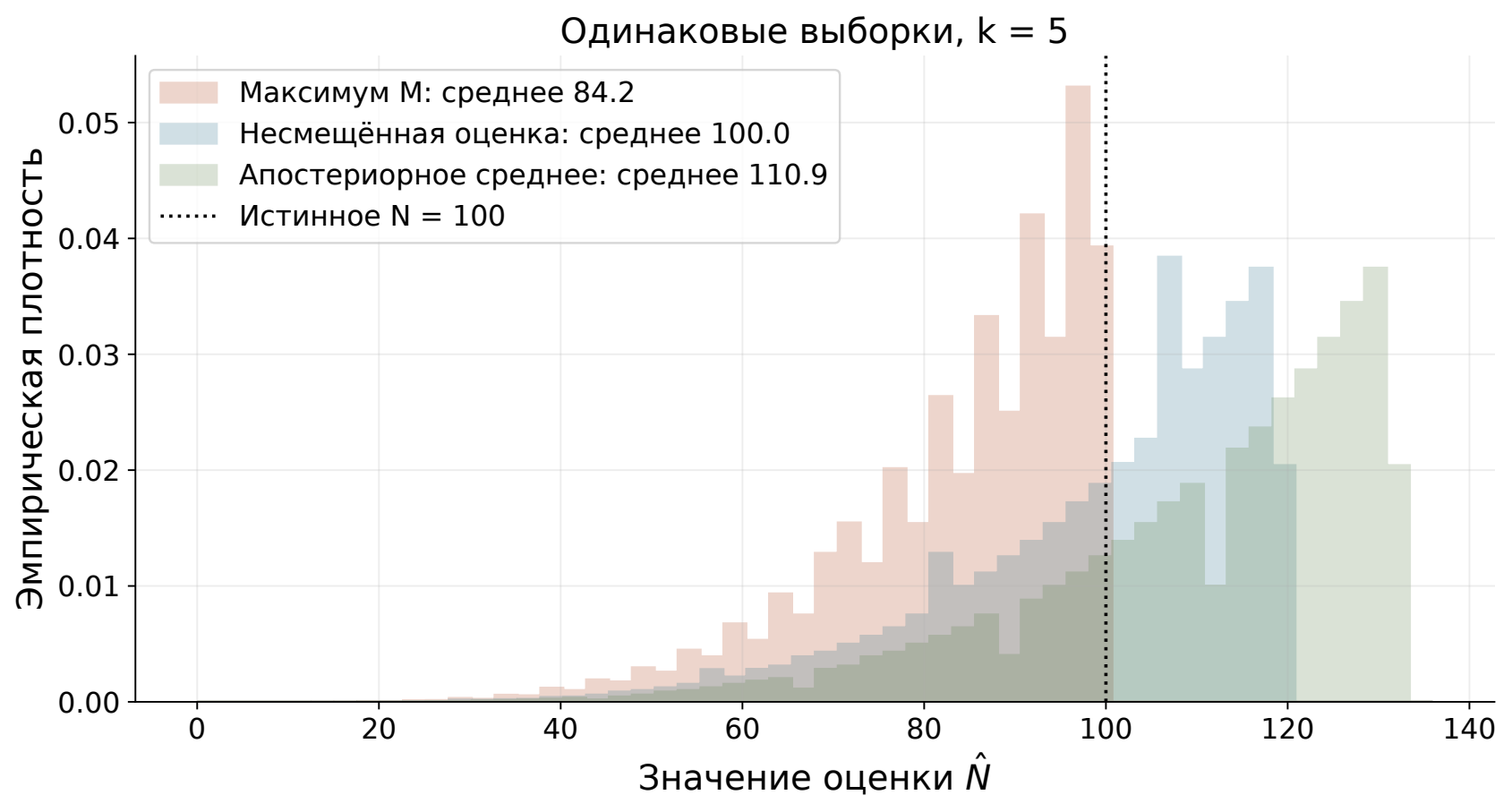
$$\hat{N} = m\left(1+\frac{1}{k}\right) - 1 = m + \left(\frac{m}{k} - 1\right).$$

- Вполне интуитивный результат: к максимуму нужно добавить «бонус», который уменьшается с ростом k

Распределение несмещённой оценки



Сравнение оценок числа танков



Как подсчитать немецкие танки

- А как по-байесовски? Мы хотим найти

$$p\left(N|M,k\right) = \frac{p\left(M|N,k\right)p\left(N|k\right)}{p\left(M|k\right)}.$$

- Мы уже подсчитали, что $p\left(M=m|N\right) = \binom{m-1}{k-1}/\binom{N}{k}$; знаменатель будем считать как нормировочную константу

$$p\left(M|k\right) = \sum_{N=M}^{\infty} p\left(M|N,k\right)p\left(N|k\right).$$

- Теперь нужно выбрать априорное распределение $p\left(N|k\right)$: какое взять? Мы бы хотели отразить в нём отсутствие информации о значении N, но равномерное распределение, как для параметра монетки, выбрать не получится, ведь нам нужно распределение на всех натуральных числах

Как подсчитать немецкие танки

- Давайте возьмём равномерное априорное распределение $p(N|k)$ на большом объемлющем интервале $k, \dots, \Omega$ (меньше k номеров при k наблюдениях получиться не может):

$$p(N|k) = \frac{1}{\Omega - k + 1} \quad \text{для всех} \quad k \le N,\ N \le \Omega.$$

- Соответственно, получается, что для $k \le N$ и $N < \Omega$

$$p(N|M,k) = \frac{p(M|N,k)\,p(N|k)}{\sum_{n=M}^{\Omega} p(M|n,k)\,p(n|k)} = \frac{\frac{1}{\Omega-k+1}p(M|N,k)}{\sum_{n=M}^{\Omega} \frac{1}{\Omega-k+1}p(M|n,k)} =$$

$$= \frac{\binom{M-1}{k-1}/\binom{N}{k}}{\sum_{n=M}^{\Omega}\binom{M-1}{k-1}/\binom{n}{k}} = \frac{\binom{N}{k}^{-1}}{\sum_{n=M}^{\Omega}\binom{n}{k}^{-1}},$$

а для других значений N эта вероятность равна нулю.

Как подсчитать немецкие танки

- Как выбрать верхний предел Ω? Можно просто заведомо большой, но технически препятствий к тому, чтобы положить $\Omega = \infty$, в формулах нет, ряд $\sum_{n=M}^{\infty} \binom{n}{k}^{-1}$ сходится для $k \geq 2$
- Это называется *некорректным априорным распределением* (improper prior): распределение «расходится», но после умножения на любое правдоподобие начинает сходиться, и байесовский вывод можно вести как обычно
- Например, бета-распределение для монетки без влияния на правдоподобие, то есть $B(0, 0)$:

$$p\left(\theta|0,0\right) \propto \frac{1}{\theta}\frac{1}{1-\theta} \quad \text{для} \quad \theta \in (0,1).$$

Как подсчитать немецкие танки

- А для танков, применив некорректное априорное распределение с $\Omega = \infty$, приходим к апостериорному распределению

$$p\left(N|M,k\right) = \frac{\binom{N}{k}^{-1}}{\sum_{n=M}^{\infty}\binom{n}{k}^{-1}}.$$

- Ряд просуммировать трудно, давайте просто поверим, что

$$\sum_{n=M}^{\infty}\binom{n}{k}^{-1} = \binom{M}{k}^{-1}\frac{M}{k-1}, \quad \text{а значит,}$$

$$p\left(N|M,k\right) = \frac{k-1}{M}\binom{M}{k}\binom{N}{k}^{-1} \quad \text{для} \quad N = M, M+1, \ldots.$$

Как подсчитать немецкие танки

- Ответ (этот результат имеет смысл только для $k > 1$):

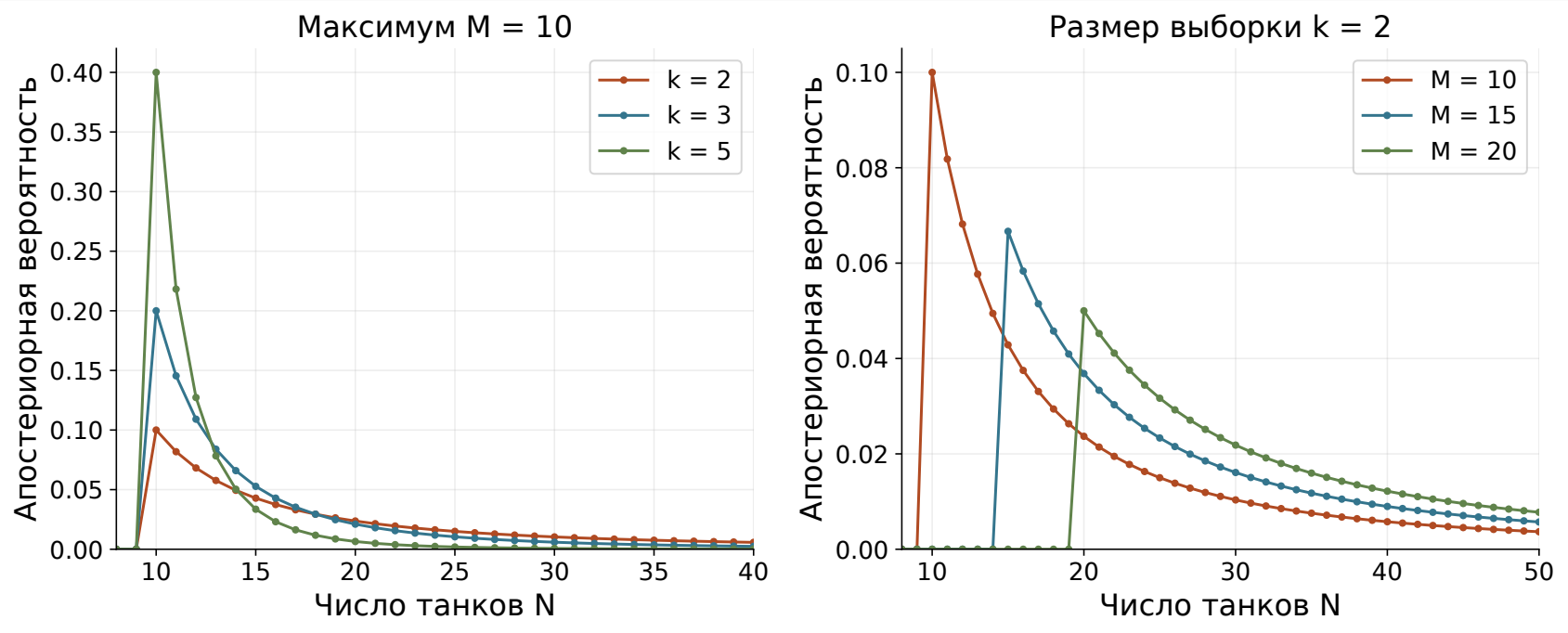
$$p\left(N|M,k\right) = \frac{k-1}{M}\binom{M}{k}\binom{N}{k}^{-1} \quad \text{для} \quad N = M, M+1, \ldots.$$

- Получается *факториальное распределение* на $N - k$, его среднее и дисперсия известны:

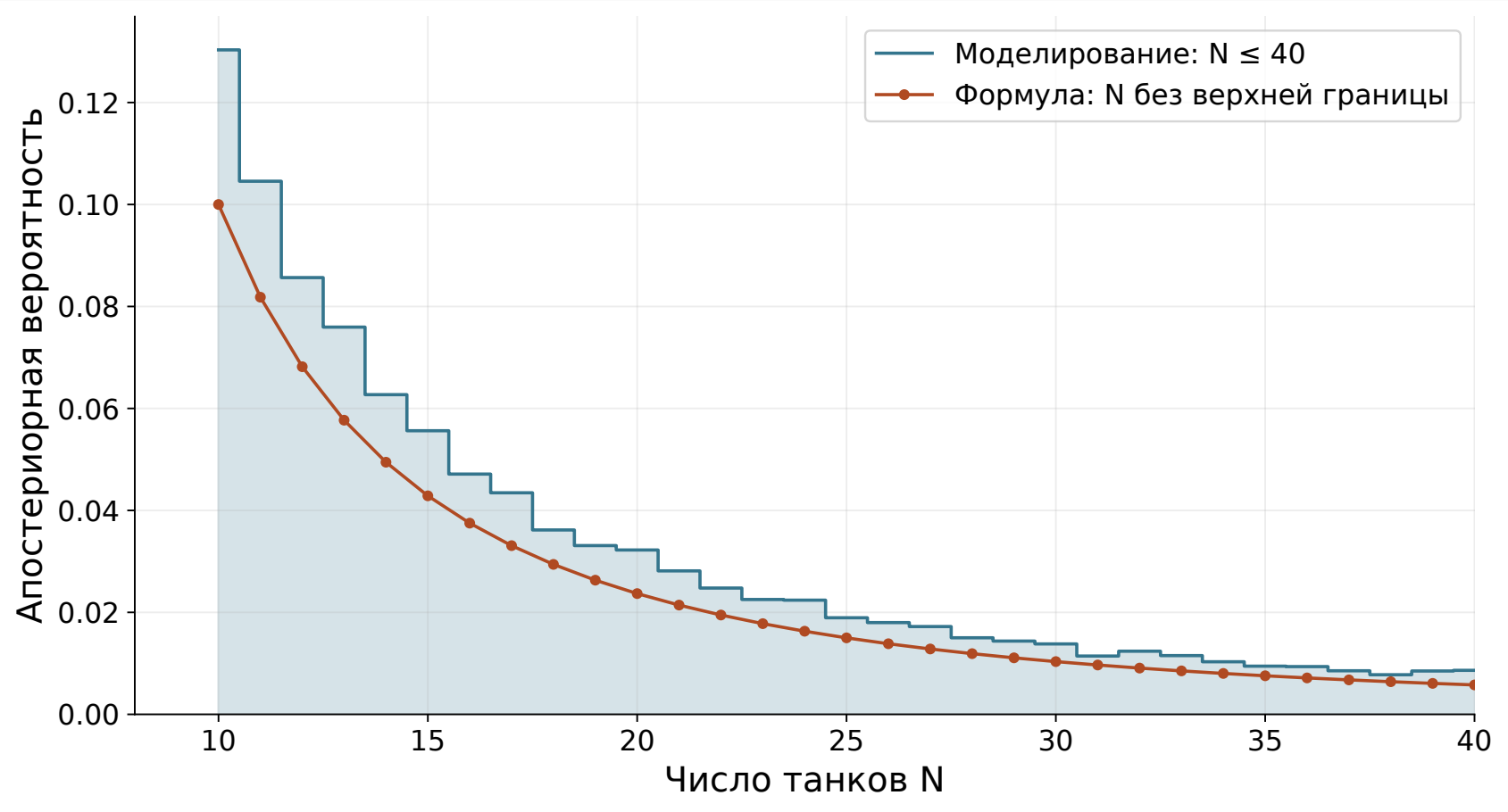
$$\mathbb{E}\left[N \mid M,k\right] = \frac{k-1}{k-2}\left(M-1\right) \quad \text{для} \quad k > 2,$$

$$\operatorname{Var}\left[N|M,k\right] = \frac{(k-1)(M-1)(M-k+1)}{(k-2)^2(k-3)} \quad \text{для} \quad k > 3.$$

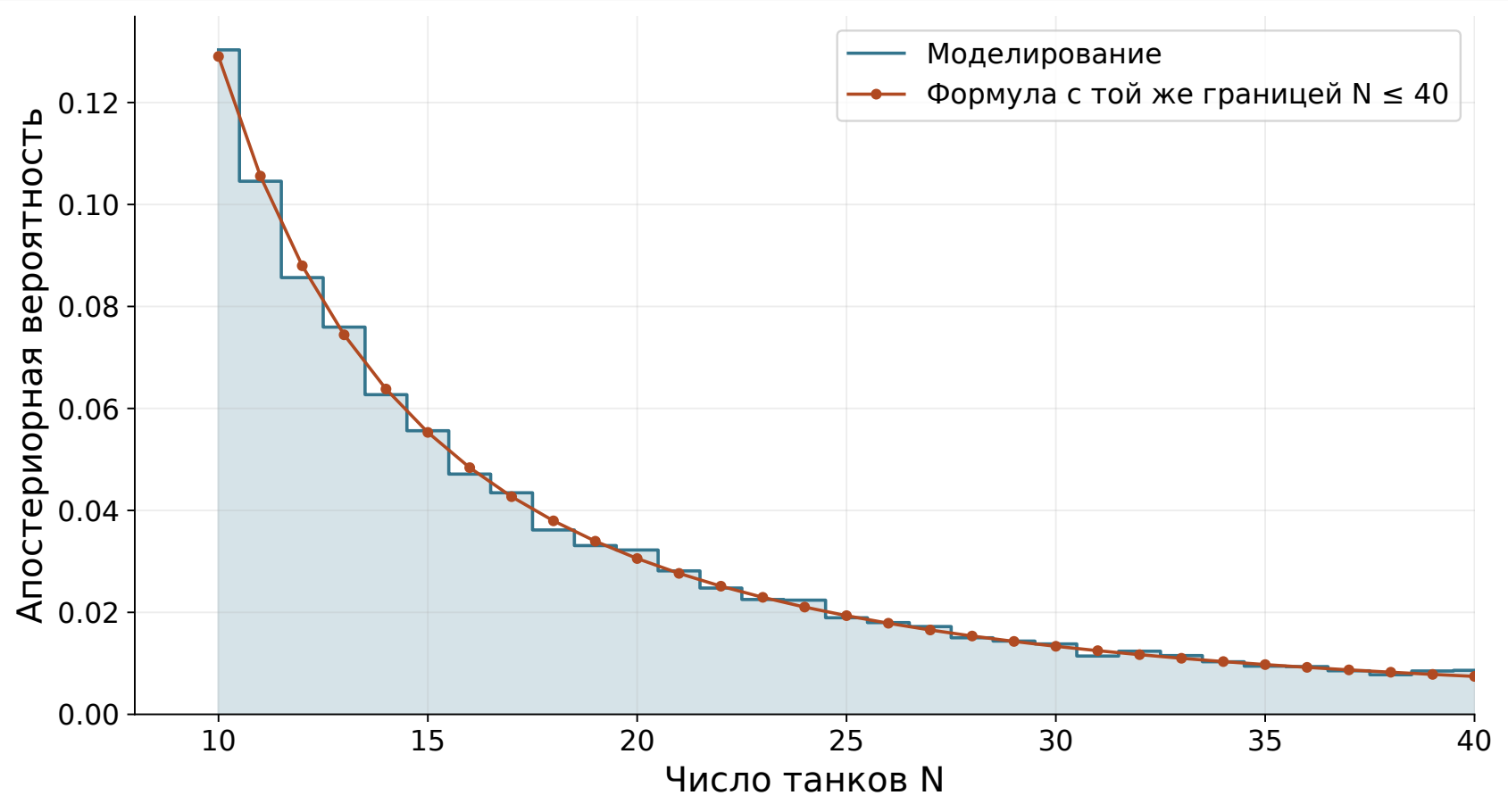
Апостериорное распределение числа танков



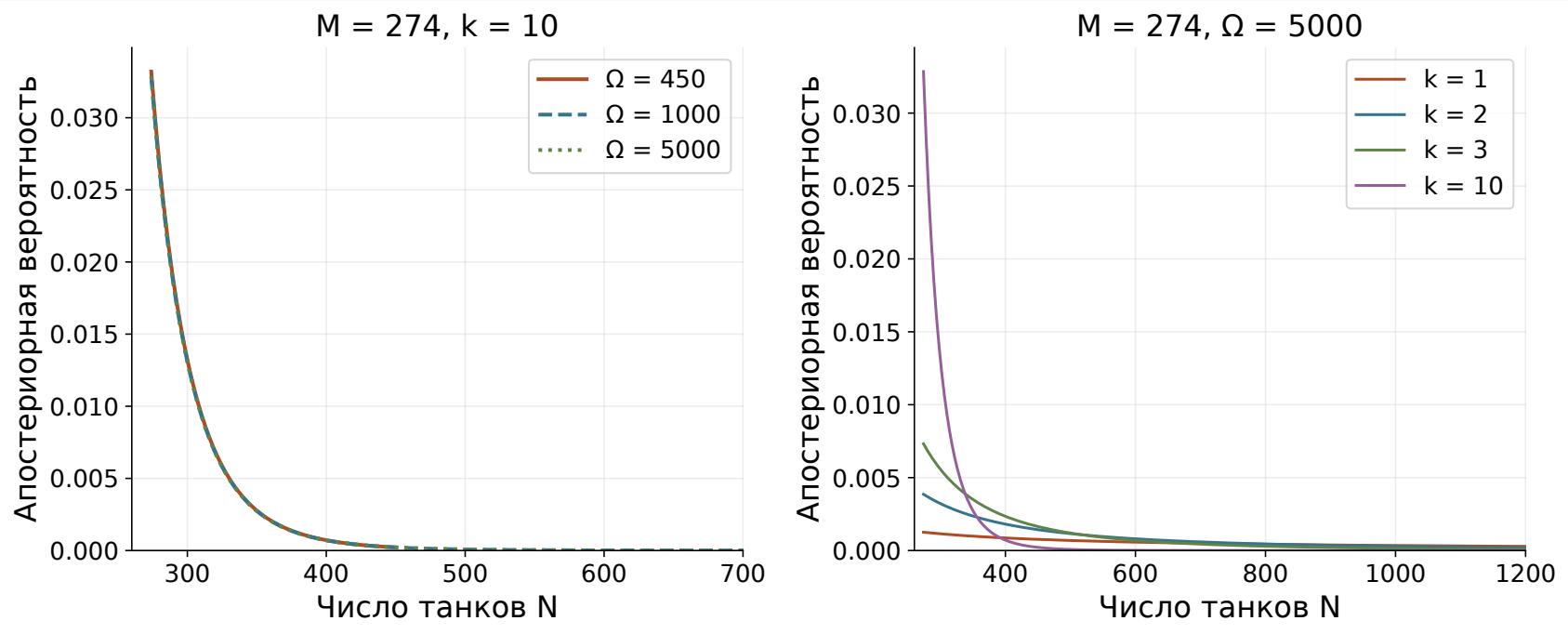
Конечное и бесконечное априорные распределения



Моделирование с одинаковой нормировкой



Верхняя граница априорного распределения



Как считать уток

- Сколько уток водится в России? Как мы можем это оценить?

Как считать уток

- Сколько уток водится в России? Как мы можем это оценить?
- Метода Линкольна — *Петерсена* (Lincoln–Petersen method, 1896 и 1930):

$$\frac{\text{Число снова пойманных}}{\text{Размер второй выборки}} = \frac{\text{Число отмеченных}}{\text{Размер выборки}},$$

то есть

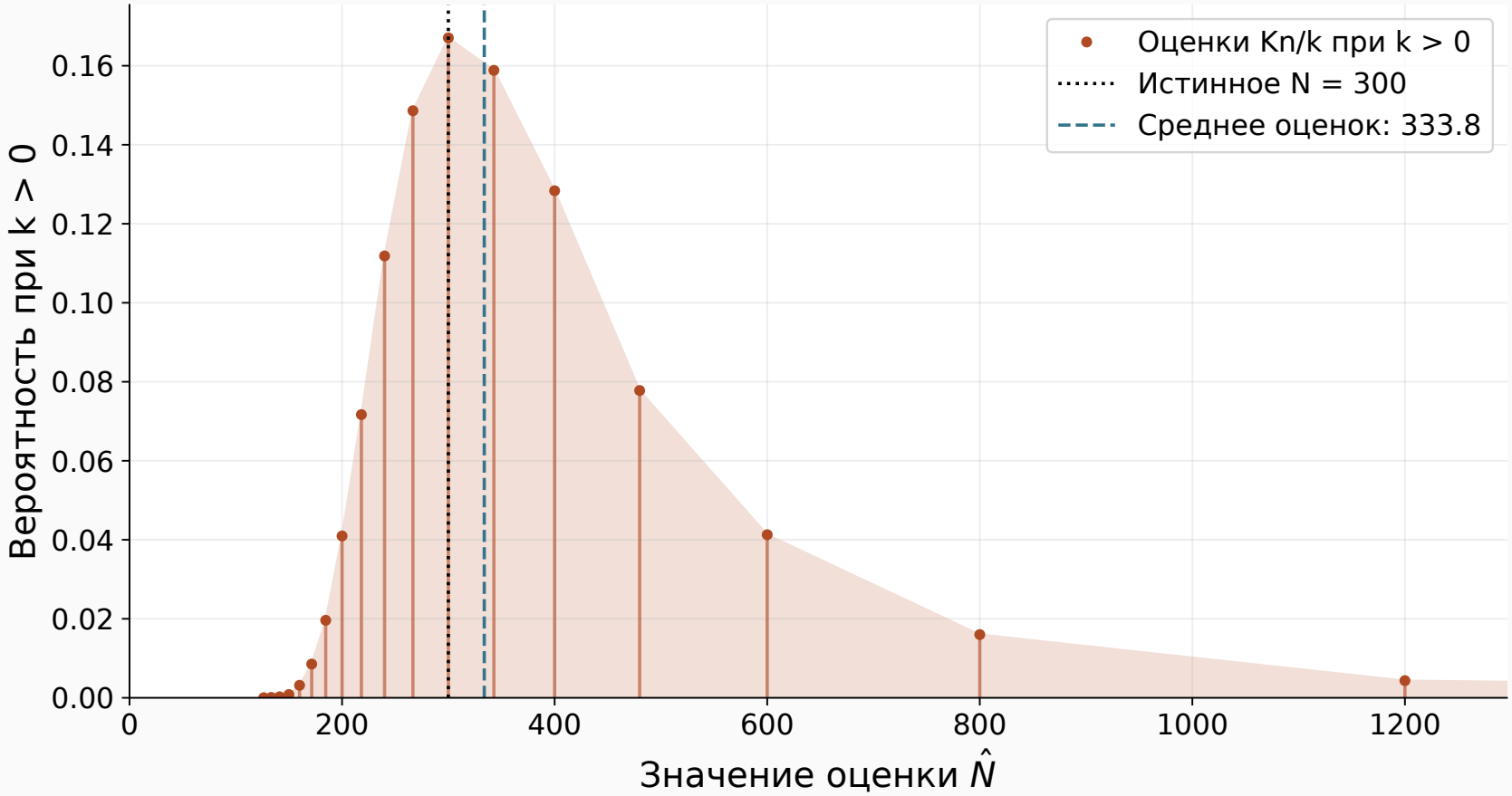
$$\text{Размер выборки} = \frac{\text{Размер второй выборки} \times \text{Число отмеченных}}{\text{Число снова пойманных}}.$$

- Кажется, тут трудно что-то добавить? Но какие здесь предположения?

Как считать уток

- Популяция предполагается *замкнутой*, т.е. птицы не рождаются и не умирают во время эксперимента
- Предполагается, что обе выборки делаются *равномерно* (на практике это противоречит первому)
- Равномерность выборки означает ещё одно предположение: даже в уже «хорошо перемешанной» совокупности ещё надо предположить, что каждая особь попадётся в наши сети с одной и той же вероятностью
- Отметки предполагаются постоянными, то есть птица не может сбросить или уничтожить кольцо
- Немало предположений, да? Но есть и математическая проблема: метод Линкольна — Петерсена несколько раз использует оценку вероятности через отношение выборок $p \approx \frac{n}{N}$; это работает, если выборки большие, но для малых n и N работать перестаёт

Распределение оценки Линкольна–Петерсена



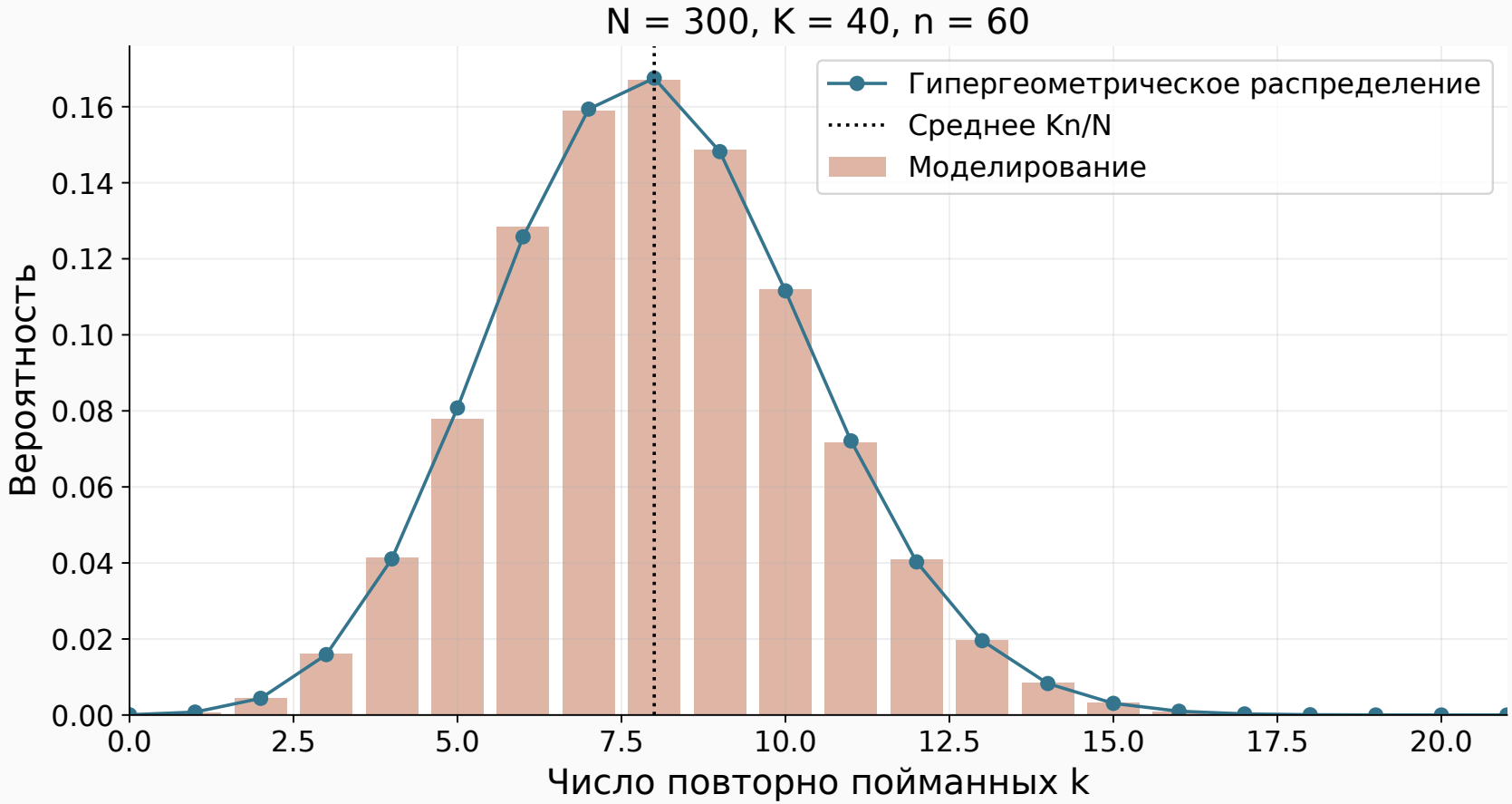
Как считать уток

- Давайте по-байесовски: правдоподобие в точности попадает в определение гипергеометрического распределения
- Нужно получить k успехов в выборке из n элементов без замещения из конечной популяции размера N с K успешными вариантами:

$$p\left(k|N,K,n\right) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

- Какое выбрать априорное распределение?

Распределение числа повторных поимок



Как считать уток

- Давайте попробуем снова воспользоваться неинформативным априорным распределением и предположить, что апостериорное распределение $p\left(N|k\right)$ будет пропорционально правдоподобию:

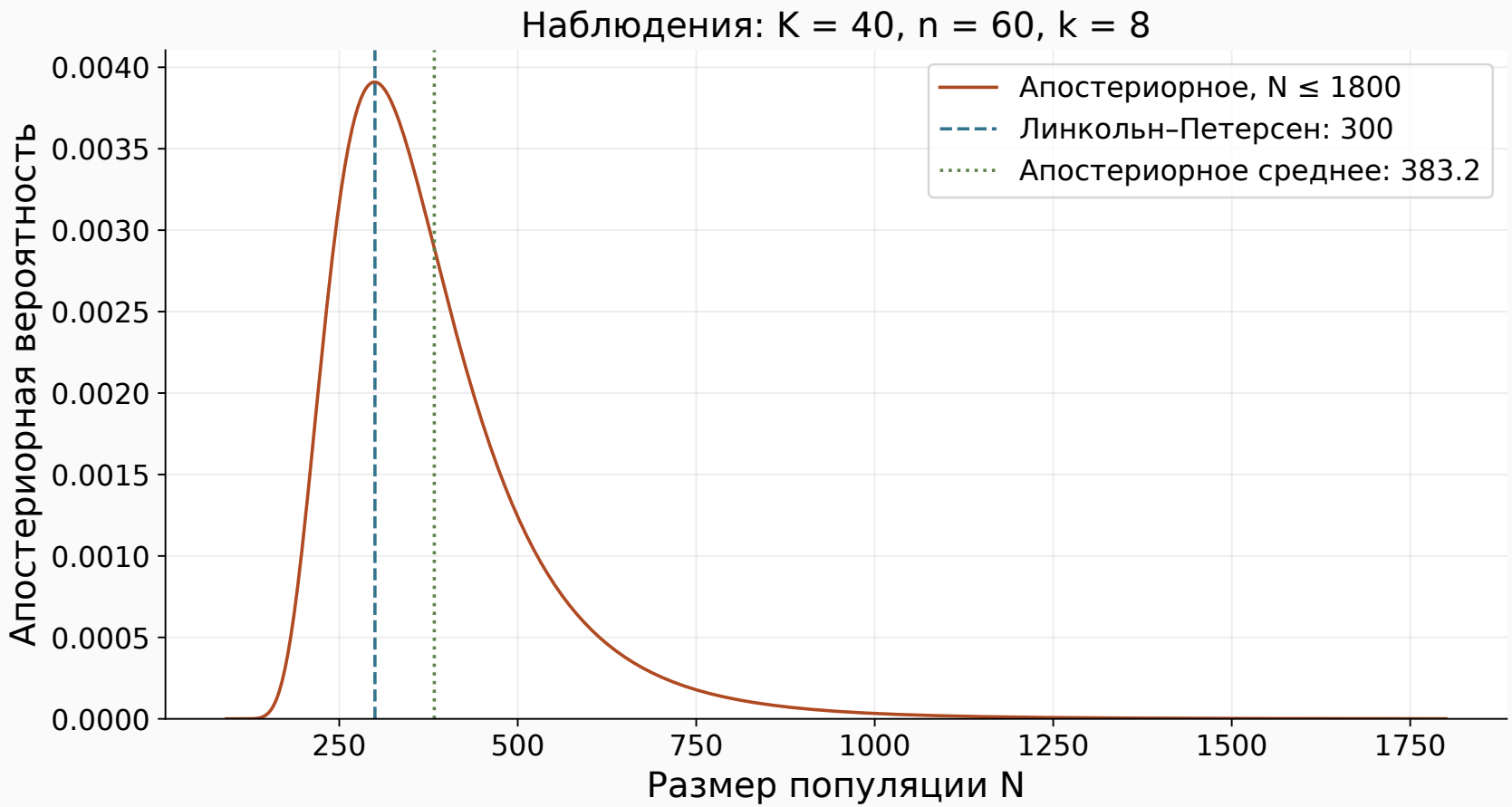
$$p\left(N|k,K,n\right) = \frac{p\left(k|N\right)}{\sum_{M=K+n-k}^{\infty} p\left(k|M\right)} = \frac{\binom{N-K}{n-k}/\binom{N}{n}}{\sum_{M=K+n-k}^{\infty}\binom{M-K}{n-k}/\binom{M}{n}}$$

- Ряд в знаменателе сходится для $k \geq 2$, и моменты апостериорного распределения можно подсчитать через гипергеометрические функции (мы не будем):

$$\mathbb{E}\left[N \mid k,K,n\right] = \frac{(K-1)(n-1)}{(k-2)} \quad \text{для} \quad k > 2,$$

$$\operatorname{Var}\left[N|k,K,n\right] = \frac{(K-1)(n-1)(K-k+1)(n-k+1)}{(k-2)^2(k-3)} \quad \text{для} \quad k > 3.$$

Апостериорное распределение числа уток

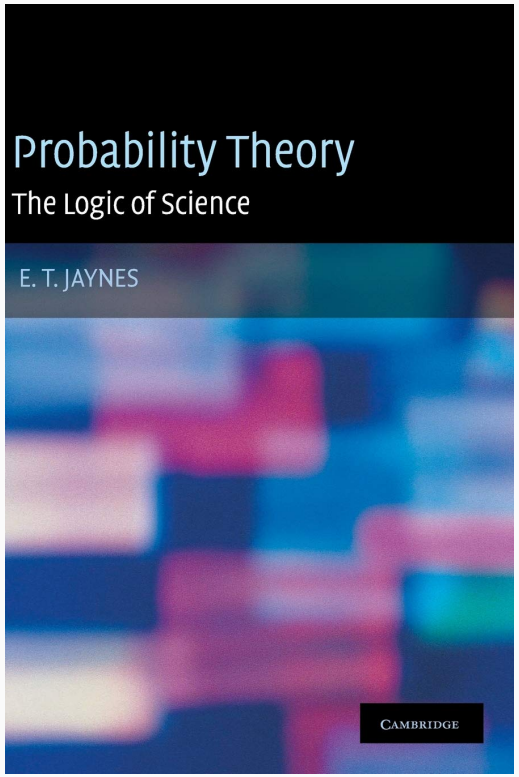


Пример Джейнса

Эдвин Джейнс

- Эдвин Томпсон Джейнс – физик, специалист по статистической механике, ввёл принцип максимума энтропии в термодинамике; развивал байесовские методы
- Probability Theory: The Logic of Science (2003):





Пример Джейнса

- Рассмотрим простую задачу байесовского вывода: предположим, что перед нами урна, в которой N шаров, R из которых красные, а $N - R$ белые
- Мы достаём из урны n шаров, r из которых оказываются красными
- Легко найти распределение пар (n, r) по заданным N и R: вероятность вынуть из урны сначала r красных шаров подряд, а затем w белых, составляет

$$\frac{R}{N}\frac{R-1}{N-1}\cdot\ldots\cdot\frac{R-r+1}{N-r+1}\cdot\frac{N-R}{N-r}\frac{N-R-1}{N-r-1}\cdot\ldots\cdot\frac{N-R-w+1}{N-r-w+1} =$$
$$= \frac{R!(N-r)!}{(R-r)!N!}\cdot\frac{(N-R)!(N-r-w)!}{(N-R-w)!(N-r)!} = \frac{R!(N-R)!(N-n)!}{(R-r)!(N-R-w)!N!}.$$

Пример Джейнса

- Эта вероятность получена для конкретной последовательности исходов, и она по симметрии не зависит от последовательности, то есть достаточно умножить её на число возможных таких последовательностей $\binom{n}{r}$:

$$p\left(r|N,R,n\right) = \frac{R!(N-R)!(N-n)!}{(R-r)!(N-R-w)!N!}\frac{n!}{r!(n-r)!} = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}},$$

  то есть гипергеометрическое распределение

Пример Джейнса

- Но теперь перед нами стоит обратная задача: мы получили данные $D = (n, r)$ и хотим что-то сказать о значениях N и R
- Байесовский вывод проходит точно так же, как выше:

$$p\left(N, R|D\right) = \frac{p\left(D|N, R\right) p\left(R|N\right) p\left(N\right)}{p\left(D\right)},$$

и здесь в знаменателе стоит нормировочная константа

$$p\left(D\right) = \sum_{N=0}^{\infty} \sum_{R=0}^{N} p\left(D|N, R\right) p\left(R|N\right) p\left(N\right),$$

а правдоподобие представляет собой гипергеометрическое распределение

Пример Джейнса

- Самый интересный вопрос: как выбрать априорное распределение $p\left(N, R\right)$? Мы его уже разложили в произведение

$$p\left(N, R\right) = p\left(R|N\right) p\left(N\right);$$

  в байесовском выводе мы выбираем априорное распределение как захотим, согласно своей интуиции, но чего мы можем захотеть от этого априорного распределения, как формализовать интуицию?
- Во-первых, есть интуиция о том, что апостериорное распределение на величину N должно зависеть от числа выбранных шаров n (как в случае танков), но не должно зависеть от того, сколько из них оказались красными: с какой стати пропорция красных шаров должна влиять на их общее число?

Пример Джейнса

- Более того в данном случае хотелось бы, чтобы полученные данные «обрезали» невозможные значения, то есть сообщали нам, что $N \geq n$, но не меняли нашу априорную интуицию о том, сколько именно шаров может быть в урне, если это число больше n. Иными словами, хотелось бы получить

$$p\left(N|D\right) = \begin{cases} c \cdot p\left(N\right), & \text{если } N \geq n, \\ 0 & \text{в противном случае,} \end{cases}$$

  где c — нормировочная константа, не зависящая от N
- Это уже нетривиальное требование! Оно накладывает серьёзные ограничения на $p\left(N, R\right)$

Пример Джейнса

- Апостериорное распределение на N равно

$$p\left(N|D\right) = \frac{p\left(N\right)p\left(D|N\right)}{p\left(D\right)} = \frac{p\left(N\right)\sum_{R=0}^{N} p\left(D|N,R\right)p\left(R|N\right)}{p\left(D\right)},$$

и требование выше значит, что

$$p\left(D|N\right) = \sum_{R=0}^{N} p\left(D|N,R\right)p\left(R|N\right) = \begin{cases} f(n,r), & \text{если } N \geq n, \\ 0 & \text{в противном случае,} \end{cases}$$

где $f(n,r)$ — это некоторая функция, зависящая от данных n и r, но не от N

Пример Джейнса

- Более того, можно вспомнить, что мы говорим о гипергеометрическом распределении: для $N \geq n$

$$\sum_{R=0}^{N} \binom{R}{r} \binom{N-R}{n-r} p\left(R|N\right) = f(n,r) \binom{N}{n}$$

- Это уже весьма нетривиальное условие на априорное распределение $p\left(R|N\right)$!
- Мы получили его из базовой интуиции о том, что данные не должны менять наше мнение об N (это в некотором смысле независимость), но мы вряд ли смогли бы записать такое условие интуитивно, без этих рассуждений
- Разумные априорные распределения на R будут удовлетворять этому условию, и байесовский вывод по N будет тривиальным: апостериорное распределение пропорционально априорному, за исключением начального отрезка $N < n$, который данные делают невозможным

Спасибо за внимание!



