Санкт-Петербургский государственный университет

Прикладная математика и информатика

Учебная практика 4 (научно-исследовательская работа)(семестр 7)

«Оценка параметров сложных распределений с применением в радиобиологии и анализе текста»

Выполнил:

Олейник М. В., группа 20.Б04-мм

Научный руководитель:

кандидат ф.-м. н., доцент

Алексеева Н. П.

Работа выполнена отлично и может быть зачтена с оценкой А

Санкт-Петербург

2023

Оглавление

Введение

Сложные распределения нашли широкое применение в описании ветвящихся процессов. Они хорошо исследованы но многие из них имеют один недостаток — перерассеяность (рассеяние больше 1), что не позволяет их применять в ситуациях, когда логарифм рассеяния имеет переменный знак.

В работе Алексеевой [1] была исследована согласованность эмпирических распределении с реинтрантно-биноминальным распределением. Эксперимент заключался в выявлении количества ядерных аномалий (таких, как ядерные протрузии, межъядерные мосты и гантелевидные ядра) в злокачественных опухолях у облучённых крыс in vitro и in vivo через 52 часа после X-облучения в дозах 5–45 Гр.

Также анализ текстов [2] показал применимость модели отрицательно-биномиального распределения к встречаемости многих слов по главам. Однако некоторые из них дают плохую согласованность из-за специфического вида эмпирических распределений.

Моя задача состоит в проверке согласованности эмпирического распределения, полученного в работе Алексеевой [1], с различными сложными распределениями, которые также следует применить в анализе встречаемости слов в тексте. Необходимо рассчитать параметры этих распределений, промоделировать, вычислить критерий согласованности с эмпирическими данными и сравнить с моделью реинтрантно-биномиального распределения для случая радиобиологии.

В подотчётном семестре мною был переработан текст первой главы для более последовательной структуры, добавлен пример моделирования биномиально-логарифмического распределения в R (стр. 14), сделан анализ логарифмически-биномиального распределения (раздел 2.2), в методе максимального правдоподобия (раздел 2.3) вместо перебора одного параметра осуществлялся поиск максимума численными методами многомерной оптимизации, в разделе 2.4 применён новый подход оценки максимального правдоподобия и добавлено логарифмически-биномиальное распределение с анализом его применимости к эмпирическим данным, начат анализ встречаемости слов в тексте с применением биномиально-логарифмического распределения в разделе 2.5.

Глава 1

Анализ сложных распределений

1.1. Определения и предпосылки

Перед началом исследования свойств и моделирования сложных распределений введём основные понятия, используемые в работе.

1.1.1. Производящая функция

В общем смысле, производящая функция — это ряд:

$$A(s) = a_0 + a_1 s + a_2 s^2 + \dots,$$

сходящийся при $-s_0 < s < s_0$, где $\{a_i\}_{i=0}^{\infty}$ — последовательность действительных чисел [3]. Соответственно, если имеется дискретное распределение

$$P(X = j) = p_j, \quad j = 0, 1, 2, \dots,$$

то

$$f(s) = p_0 + p_1 s + p_2 s^2 + \dots$$

будет его производящей функцией. Например, производящей функцией биномиального распределения:

$$B(s) = \sum_{k=0}^{n} C_n^k (ps)^k q^{n-k} = (q + sp)^n.$$

Для нахождения характеристик сложных произведений, а также для оценки вероятностей нам понадобятся свойства производящей функции [3] (ξ — случайная дискретная величина, $h(t)$ — её производящая функция):

1. $\mathsf{E}\xi = h'(1)$;

2. $\mathsf{D}\xi = h''(1) + h'(1) - (h'(1))^2$;

3. $P(\xi = k) = \frac{h^{(k)}(0)}{k!}$.

1.1.2. Реинтрантно-биномиальное распределение

Реинтрантно-биномиальное распределение, используемое в работе [1], имеет производящую функцию вида:

$$f(v) = (p_0(p_1 v + q_1)^{n_1} + q_0)^{n_0},$$

то есть, как видно из структуры, это суперпозиция производящих функций двух биномиальных распределений.

Также в статье представлены формулы вычисления математического ожидания и дисперсии:

$$\mathsf{E}X = n_0 n_1 p_0 p_1,$$
$$\mathsf{D}X = n_0 n_1 p_0 p_1 (n_1 p_1 (1 - p_0) - p_1 + 1).$$

Это распределение хорошо описывает экспериментальные данные, однако имеет целых четыре параметра, которые при интерпретации и нахождении оценок максимального правдоподобия сводятся к двум произведениям реинтрантных компонент, что говорит о возможном упрощении его структуры.

1.2. Логарифмическое распределение

Единственное (из часто используемых) дискретное распределение имеющее — без составления в сложные распределения — переменный знак логарифма рассеяния — это логарифмическое распределение ($\xi \sim Log(q)$) с такими вероятностями:

$$P(\xi = k) = \frac{-1}{\ln(1-p)} \frac{p^k}{k}, \quad p \in (0,1), k = 1, 2, \dots.$$

Оно имеет производящую функцию:

$$h(t) = \frac{\ln(1-pt)}{\ln(1-p)}.$$

То есть вероятности — это коэффициенты ряда Тейлора разложения логарифма с нормирующим множителем. Вычислим математическое ожидание, используя свойства производящей функции:

$$\mathsf{E}\xi = h'(1) = \left(\frac{-p}{\ln(1-p)\cdot(1-pt)}\right)(1) = \frac{-p}{\ln(1-p)\cdot(1-p)},$$

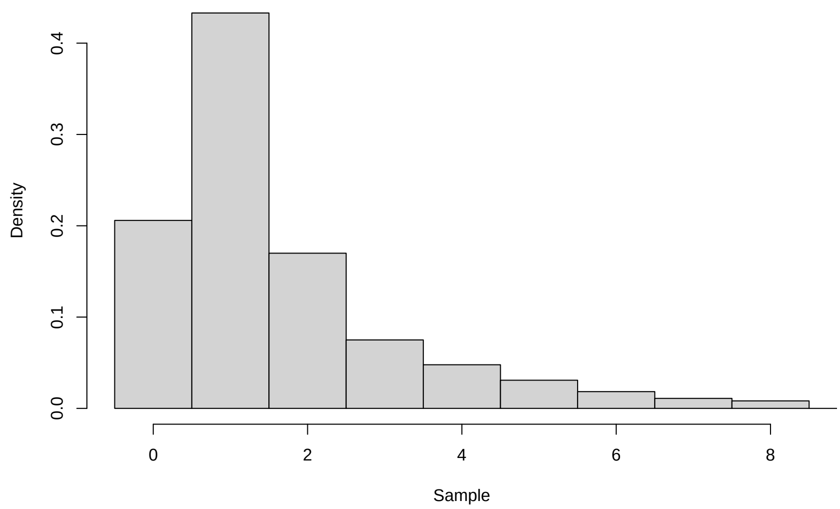

Рис. 1.1. Промоделированное логарифмическое распределение с параметром. $p_0 = 0.2$ и $p = 0.75$

и дисперсию:

$$h''(t) = \frac{-p^2}{\ln(1-p)\cdot(1-pt)^2}$$

$$\mathsf{D}\xi = h''(1) + h'(1) - \left(h'(1)\right)^2 = \frac{-p^2}{\ln(1-p)\cdot(1-p)^2} + \frac{-p}{\ln(1-p)\cdot(1-p)} - \frac{p^2}{\ln^2(1-p)\cdot(1-p)^2} = -p\cdot\frac{p+\ln(1-p)}{\ln^2(1-p)\cdot(1-p)^2}.$$

Тогда рассеяние равно:

$$e\xi = \frac{\mathsf{D}\xi}{\mathsf{E}\xi} = \frac{p+\ln(1-p)}{\ln(1-p)\cdot(1-p)}.$$

Найдём при каком p логарифм рассеяния меняет знак:

$$e\xi = \frac{p+\ln(1-p)}{\ln(1-p)\cdot(1-p)} = 1$$

$$p+\ln(1-p) = \ln(1-p) - p\cdot\ln(1-p)$$

$$p(1+\ln(1-p)) = 0, p \neq 0$$

$$\ln(1-p) = -1$$

$$1-p = e^{-1}$$

$$p = 1-e^{-1}.$$

То есть логарифм рассеяния при $p \in (0,1)$ может менять знак.

Однако данная дискретная случайная величина может принимать значения только начиная с 1, а задача требует, чтобы она могла принимать значение 0. Для этого введём дополнительный параметр p_0:

$$\begin{aligned}
&P(\xi = 0) = p_0;\\
&P(\xi = k) = (1 - p_0) \cdot \frac{-1}{\ln(1-p)} \frac{p^k}{k}, \quad p \in (0,1), k = 1, 2, \ldots.
\end{aligned}$$

Производящая функция этого распределения:

$$g(t) = (1 - p_0) \cdot \frac{\ln(1-pt)}{\ln(1-p)} + p_0,$$

математическое ожидание:

$$\mathsf{E}\xi = g'(1) = (1 - p_0) \cdot \frac{-p}{\ln(1-p) \cdot (1-p)},$$

дисперсия:

$$\begin{aligned}
h''(t) =&(1 - p_0) \cdot \frac{-p^2}{\ln(1-p) \cdot (1-pt)^2}\\
\mathsf{D}\xi =&h''(1) + h'(1) - \left(h'(1)\right)^2 = (1 - p_0) \cdot \frac{-p^2}{\ln(1-p) \cdot (1-p)^2} + (1 - p_0)\cdot\\
&\cdot \frac{-p}{\ln(1-p) \cdot (1-p)} - (1 - p_0)^2 \cdot \frac{p^2}{\ln^2(1-p) \cdot (1-p)^2} =\\
=& - p(1 - p_0) \cdot \frac{(1-p_0)p + \ln(1-p)}{\ln^2(1-p) \cdot (1-p)^2},
\end{aligned}$$

рассеяние:

$$e\xi = \frac{\mathsf{D}\xi}{\mathsf{E}\xi} = \frac{p(1-p_0) + \ln(1-p)}{\ln(1-p) \cdot (1-p)}.$$

Аналогично предыдущим рассуждениям приходим к тому, что смена знака логарифма рассеяния происходит при:

$$p = 1 - e^{p_0 - 1}.$$

Промоделированное логарифмическое распределение с параметром представлено на рис. 1.1.

Если мы посмотрим на данные в статье [1], то убедимся, что их мода не всегда равна 0 или 1, а, значит, и это распределение нам не подходит, но не из-за пере- или недорассеяности, а по причине формы, так как оно является простым разложением монотонного логарифма, то есть его мода всегда в 1 (при большом p_0 — в 0).

1.3. Рассеяние случайной величины

В статье Алексеевой [1] показано, что эмпирические данные должны иметь модель распределения, логарифм рассеяния которого имеет переменный знак, или рассеяние может быть больше и меньше 1, что то же самое. Однако многие дискретные распределения (биномиальное, пуассоновское, геометрическое) не обладают таким свойством. Было выдвинуто предположение, что логарифмическое распределение:

$$P(\xi = k) = \frac{-1}{\ln(1-p)} \frac{p^k}{k}, \quad k = 1, 2, \dots,$$

с параметром p_0, определяющим вероятность в 0, окажется верным, однако из-за того, что данное распределение является разложение логарифма в ряд, то его максимум всегда будет в 0 или 1 — это не соответствовало эмпирическим данным.

Хотелось бы понять, как ведёт себя рассеяние случайной величины, удовлетворяющей сложному закону распределения. Для этого докажем следующую лемму.

Лемма 1. *Пусть S_N — сумма случайного числа N случайных величин ξ_i, одинаково распределённых и не зависимых между собой и N. Тогда для её рассеяния справедлива формула:*

$$eS_N = \mathsf{E}\xi eN + e\xi.$$

Доказательство. Докажем это утверждение через производящие функции случайных величин. Пусть $g(t)$ — производящая функция случайной величины ξ (ξ имеет такое же распределение, как и ξ_i), а $h(t)$ — производящая функция случайной величины N. Тогда производящая функция случайной величины S_N — $h(g(t))$. Отсюда следует математическое ожидание (по свойствам производящих функций):

$$\mathsf{E}S_N = (h(g))'(1) = h'(g(1))g'(1) = \mathsf{E}\xi\mathsf{E}N,$$

и дисперсия:

$$\mathsf{D}S_N = (h(g))''(1) + (h(g))'(1) - \left((h(g))'(1)\right)^2 = (\mathsf{E}\xi)^2\,\mathsf{D}N + \mathsf{E}N\mathsf{D}\xi.$$

Поделив дисперсию на математическое ожидание как раз получим рассеяние:

$$eS_N = \frac{\mathsf{D}S_N}{\mathsf{E}S_N} = \frac{(\mathsf{E}\xi)^2\,\mathsf{D}N + \mathsf{E}N\mathsf{D}\xi}{\mathsf{E}\xi\mathsf{E}N} = \mathsf{E}\xi eN + e\xi.$$

□

Из леммы 1 следует, что, так как мы рассматриваем дискретные распределения с неотрицательными значениями случайной величины, то, если рассеяние у слагаемых случайных величин больше 1, мы получим рассеяние всей суммы больше 1. Поэтому для построения модели, согласованной с эмпирическими данными, мы возьмём суперпозицию распределений, внутреннее из которых имеет рассеяние хотя бы при каких-то значениях параметров меньше 1.

Глава 2

Сложные распределения на основе биномиального и логарифмического

2.1. Биномиально-логарифмическое распределение

В прошлой главе была приведена лемма 1, из которой следовало, что для переменности знака логарифма рассеяния сложного распределения необходима переменность знака логарифма рассеяния у слагаемых случайной суммы случайных величин. Тогда в качестве такого распределения рассмотрим знакомое по прошлой главе логарифмическое распределение, а количество слагаемых будет подчинятся биномиальному закону. Тогда производящая функция такой суперпозиции имеет вид:

$$h(t) = \left(p\frac{\ln(1-qt)}{\ln(1-q)} + 1 - p\right)^n.$$

Определение 1. *Такую суперпозицию будем называть биномиально-логарифмическим распределением.*

Вычислим основные характеристики этого распределения через характеристики образующих его распределений и свойства производящих функций:

$$\mathsf{E}S_N = \mathsf{E}\xi\mathsf{E}N = \frac{-npq}{\ln(1-q)(1-q)},$$

$$\mathsf{D}S_N = (\mathsf{E}\xi)^2\,\mathsf{D}N + \mathsf{E}N\mathsf{D}\xi = \frac{-np^2q^2 - npq\ln(1-q)}{\ln^2(1-q)(1-q)^2},$$

$$eS_N = \mathsf{E}\xi eN + e\xi = \frac{qp + \ln(1-q)}{\ln(1-q)(1-q)}.$$

Итого, при $p = -\ln(1-q)$ логарифм рассеяние меняет знак.

Для оценки параметров по методу максимального правдоподобия необходимо вычислить теоретические вероятности биномиально-логарифмического распределения.

2.1.1. Вероятности

Как известно из свойств производящей функции [3]: если $h(t)$ — производящая функция дискретного распределения, то вероятности этого произведения равны:

$$P(\xi = k) = \frac{h^{(k)}(0)}{k!}.$$

Тогда для производящей функции сложного распределения также справедливо:

$$P(S_N = k) = \frac{\left(h(g)\right)^{(k)}(0)}{k!},$$

где $h(t)$ — производящая функция N, $g(t)$ — производящая функция ξ.

Воспользуемся формулой Фаа-ди-Бруно для производных сложной функции:

$$\frac{d^n}{dx^n} h(g(x)) = \sum \frac{n!}{m_1! m_2! \cdots m_n!} h^{(m_1+m_2+\cdots+m_n)}(g(x)) \cdot \prod_{j=1}^{n} \left(\frac{g^{(j)}(x)}{j!} \right)^{m_j},$$

где сумма идёт по всем кортежам $(m_1, m_2, \ldots, m_n)$ длины n из неотрицательных чисел удовлетворяющих ограничению:

$$1 \cdot m_1 + 2 \cdot m_2 + \cdots + n \cdot m_n = n.$$

Причём в случае биномиально-логарифмического распределения, так как:

$$g(0) = \frac{\ln(1 - q \cdot 0)}{\ln(1-q)} = 0,$$

то

$$h^{(k)}(0) = k! \cdot P(N = k), \quad \forall k = 0, 1, 2, \ldots$$

Тогда формула приобретает вид:

$$\begin{aligned} P(S_N = k) = \left(\frac{d^k}{dx^k} h(g) \right)(0) \cdot \frac{1}{k!} = \sum \frac{1}{m_1! m_2! \cdots m_k!} h^{(m_1+m_2+\cdots+m_k)}(0) \cdot \\ \cdot \prod_{j=1}^{k} \left(\frac{g^{(j)}(0)}{j!} \right)^{m_j} = \\ = \sum \frac{1}{m_1! m_2! \cdots m_k!} (m_1 + m_2 + \cdots + m_k)! \cdot P(N = m_1 + m_2 + \cdots + m_k) \cdot \\ \cdot \prod_{j=1}^{k} \left(P(\xi = j) \right)^{m_j}. \end{aligned}$$

Посчитанные вероятности для $k = 0, 1, 2, 3, 4$:

$$\begin{aligned}
0!P(S_N = 0) =& P(N = 0)\\
1!P(S_N = 1) =& P(N = 1) \cdot P(\xi = 1)\\
2!P(S_N = 2) =& 2!P(N = 2) \cdot (P(\xi = 1))^2 + P(N = 1) \cdot 2!P(\xi = 2)\\
3!P(S_N = 3) =& 3!P(N = 3) \cdot (P(\xi = 1))^3 + 3 \cdot 2!P(N = 2) \cdot P(\xi = 1) \cdot 2!P(\xi = 2)+\\
&+ P(N = 1) \cdot 3!P(\xi = 3)\\
4!P(S_N = 4) =& 4!P(N = 4) \cdot (P(\xi = 1))^4 + 6 \cdot 3!P(N = 3) \cdot (P(\xi = 1))^2 \cdot 2!P(\xi = 2)+\\
&+ 3 \cdot 2!P(N = 2) \cdot (2!P(\xi = 2))^2 + 4 \cdot 2!P(N = 2) \cdot P(\xi = 1) \cdot 3!P(\xi = 3)+\\
&+ P(N = 1) \cdot 4!P(\xi = 4).
\end{aligned}$$

Преобразуем, подставив вероятности соответствующих распределений:

$$\begin{aligned}
P(S_N = 0) =& \frac{1}{0!}(1-p)^n\\
P(S_N = 1) =& \frac{1}{1!}np(1-p)^{n-1} \cdot \frac{-q}{\ln(1-q)}\\
P(S_N = 2) =& \frac{1}{2!}np(1-p)^{n-2} \cdot \frac{q^2}{\ln(1-q)}\left((n-1)p\frac{1}{\ln(1-q)} - (1-p)\right)\\
P(S_N = 3) =& \frac{1}{3!}np(1-p)^{n-3} \cdot \frac{-q^3}{\ln(1-q)}\Bigg((n-1)(n-2)p^2\frac{1}{\ln^2(1-q)} -\\
&- 3(n-1)p(1-p)\frac{1}{\ln(1-q)} + 2(1-p)^2\Bigg)\\
P(S_N = 4) =& \frac{1}{4!}np(1-p)^{n-4} \cdot \frac{q^4}{\ln(1-q)}\Bigg((n-1)(n-2)(n-3)p^3\frac{1}{\ln^3(1-q)} -\\
&- 6(n-1)(n-2)p^2(1-p)\cdot\\
&\cdot \frac{1}{\ln^2(1-q)} + 11p(1-p)^2\frac{1}{\ln(1-q)} - 6(1-p)^3\Bigg).
\end{aligned}$$

Тогда общая формула принимает вид:

$$P(S_N = k) = \frac{1}{k!}(1-p)^{n-k} \cdot q^k \sum_{j=0}^{k} \frac{n!}{(n-j)!} c(k, j)(p\alpha)^j (1-p)^{(k-j)},$$

где $\alpha = \frac{-1}{\ln(1-q)}$, $c(k, j)$ — число Стирлинга первого рода без знака [4], задающее число перестановок из k элементов с j циклами. Также они могут быть определены, как коэффициенты при полиноме такого вида:

$$\prod_{i=0}^{k}(i + x) = \sum_{j=1}^{k} s(k, j)x^j.$$

Например, при $k = 4$ получим

$$x(x+1)(x+2)(x+3) = 6x + 11x^2 + 6x^3 + x^4,$$

откуда $s(4,0) = 0, s(4,1) = 6, s(4,2) = 11, s(4,3) = 6, s(4,4) = 1$. Можно записать такую рекурентную формулу

$$s(k+1, j) = s(k, j-1) + ks(k, j),$$

по которой получим последовательность коэффициентов:

$k\backslash j$	0	1	2	3	4	5	6
0	1						
1	0	1					
2	0	1	1				
3	0	2	3	1			
4	0	6	11	6	1		
5	0	24	50	35	10	1	
6	0	120	274	225	85	15	1

Но нам будет достаточно первых пяти вероятностей для оценок распределения, так как полученные экспериментальные данные дают максимальное количество аномалий, равное 5.

2.1.2. Моделирование сложных распределений в R

Пусть $\{X_j\}$ — последовательность одинаково распределённых случайных величин, тогда рассмотрим

$$S_N = X_1 + X_2 + \cdots + X_N,$$

где N — случайная величина, не зависящая от X_j. Производящая функция распределения $P\{S_N = j\}$ — это суперпозиция производящих функций распределений $P\{N = n\}$ и $P\{X_j = x\}$ [3, с. 291].

Чтобы смоделировать выборку из n элементов сложного распределения, нужно смоделировать выборку из n элементов первого простого распределения, для каждого

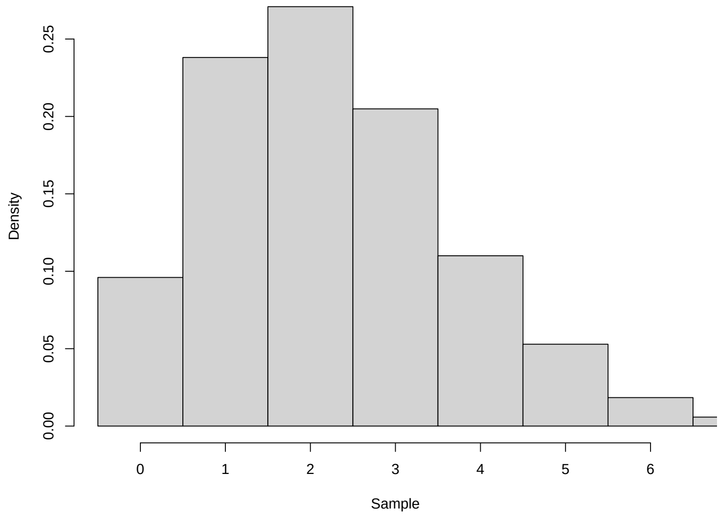

Рис. 2.1. Промоделированное биномиально-логарифмическое распределение. $q = 0.2, p = 0.25$ и $n = 8$

из которых вычисляется сумма выборки из элементов второго, чьё количество равно значению элемента из первого.

Функция, позволяющая моделировать биномиально-логарифмическое распределение в R:

```
# Modeling of binomial-logarithmic distribution.
# sumlog - accumulated probabilities of the logarithmic
# distribution.
rbinomlog <- function(num = 1, sumlog, n = 1, p = 0.5) {
        res <- c()

        for(i in rbinom(num, n, p)){
                res <- c(res, sum(0, rnlog(i, sumlog)))
        }

        return(res)
}
```

Промоделированное биномиально-логарифмическое распределение с рассеянием меньше 1 представлено на рис. 2.1.

2.2. Логарифмически-биномиальное распределение

Мы рассматривали суперпозицию биномиального и логарифмического произведения, потому что логарифм рассеяния логарифмического распределения имеет переменный знак, однако у биномиального рассеяние всегда меньше нуля, что не мешает рассматривать его на роль слагаемых в случайной сумме. Поэтому теперь обратим внимание на суперпозицию логарифмического и биномиального распределений (в обратном порядке). Производящая функция такого распределения:

$$g(t) = \frac{\ln(1 - q(pt + 1 - p)^n)}{\ln(1 - q)}.$$

Определение 2. *Такую суперпозицию будем называть логарифмически-биномиальным распределением.*

Аналогично найдём для него основные характеристики ($\xi \sim \mathrm{Bin}(n, p), N \sim \mathrm{Log}(q)$):

$$\begin{aligned}
\mathsf{E}S_N &= \mathsf{E}\xi\mathsf{E}N = \frac{-npq}{\ln(1-q)(1-q)},\\
\mathsf{D}S_N &= (\mathsf{E}\xi)^2\,\mathsf{D}N + \mathsf{E}N\mathsf{D}\xi = npq\frac{-npq - np\ln(1-q) - (1-p)(1-q)\ln(1-q)}{\ln^2(1-q)(1-q)^2},\\
eS_N &= \mathsf{E}\xi eN + e\xi = \frac{npq + np\ln(1-q) + (1-p)(1-q)\ln(1-q)}{\ln(1-q)(1-q)}.
\end{aligned}$$

Для этого распределения знак логарифма рассеяния не зависит от параметра p, однако зависит от n. Знак меняется при $n = \frac{(1-q)\ln(1-q)}{q+\ln(1-q)}$.

2.2.1. Вероятности

Аналогично биномиально-логарифмическому распределению получим вероятности логарифмически-биномиального, используя свойства производящей функции:

$$P(S_N = k) = \frac{1}{k!}h^{(k)}(0), \quad \forall k = 0, 1, 2, \dots$$

Причём, если обозначить отдельно производящие функции для ξ и N как h и f, то получим такие формулы для производных:

$$\begin{aligned}
g(t) =& f(h(t))\\
g'(t) =& f'(h(t))h'(t)\\
g''(t) =& f''(h(t))(h'(t))^2 + f'(h(t))h''(t)\\
g'''(t) =& f'''(h(t))(h'(t))^3 + f''(h(t))2h'(t)h''(t) +\\
& + f''(h(t))h'(t)h''(t) + f'(h(t))h'''(t)\\
g^{IV}(t) =& f^{IV}(h(t))(h'(t))^4 + 6f'''(h(t))(h'(t))^2h''(t) + 3f''(h(t))(h''(t))^2 +\\
& + 4f''(h(t))h'(t)h'''(t) + f'(h(t))h^{IV}(t)
\end{aligned}.$$

Тогда итоговые вероятности будут равны:

$$\begin{aligned}
0!P(S_N = 0) =& \frac{\ln(1 - q(1-p)^n)}{\ln(1-q)}\\
1!P(S_N = 1) =& \frac{-qpn(1-p)^{n-1}}{(1 - q(1-p)^n)\ln(1-q)}\\
2!P(S_N = 2) =& \frac{-qp^2n(1-p)^{n-2}}{(1 - q(1-p)^n)\ln(1-q)}\left(\frac{qn(1-p)^n}{(1-q(1-p)^n)} + (n-1)\right)\\
3!P(S_N = 3) =& \frac{-qp^3n(1-p)^{n-3}}{(1 - q(1-p)^n)\ln(1-q)}\left(\frac{2q^2n^2(1-p)^{2n}}{(1-q(1-p)^n)^2} + \frac{3qn(n-1)(1-p)^n}{(1-q(1-p)^n)}\right.\\
& \left. + (n-1)(n-2)\right)\\
4!P(S_N = 4) =& \frac{-qp^4n(1-p)^{n-4}}{(1 - q(1-p)^n)\ln(1-q)}\left(\frac{3q^3n^3(1-p)^{3n}}{(1-q(1-p)^n)^3} + \frac{12q^2n^2(n-1)(1-p)^{2n}}{(1-q(1-p)^n)^2}\right.\\
& \left. + \frac{q(1-p)^n}{(1-q(1-p)^n)}\big(3n(n-1)^2 + 4n(n-1)(n-2)\big) + (n-1)(n-2)(n-3)\right).
\end{aligned}$$

Для определения согласованности с эмпирическим распределением нам хватит этого набора вероятностей, поэтому искать общую формулу не будем.

2.3. Метод максимального правдоподобия

Для нахождения параметров биномиально-логарифмического распределения можем применить метод максимального правдоподобия в общем случае (то есть для любых вероятностей), а для логарифмически-биномиального будем считать значения боль-

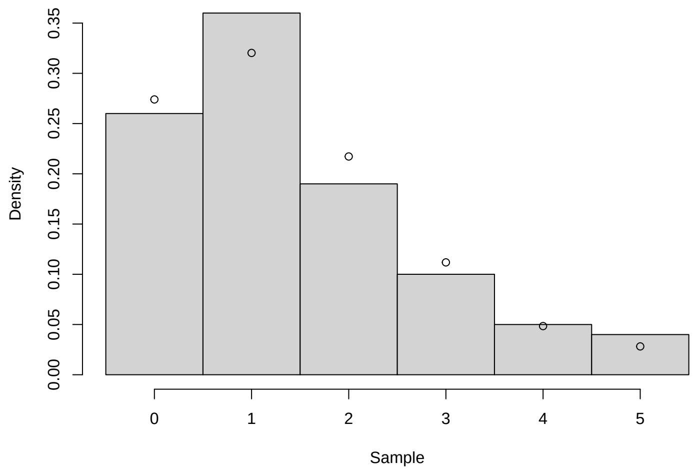

Рис. 2.2. Эмпирические частоты (столбики) и теоретические частоты (точки), вычисленные по методу максимального правдоподобия

ше 4 одинаковыми. Составим функцию правдоподобия:

$$\mathcal{L}(x_1,\dots,x_m,p,q,n)=\prod_{t=1}^{m}P(S_N=x_t,p)=$$
$$=\prod_{t=1}^{m}\frac{1}{x_t!}(1-p)^{n-x_t}\cdot q^{x_t}\sum_{j=1}^{x_t}\frac{n!}{(n-j)!}c(x_t,j)(p\alpha)^j(1-p)^{(x_t-j)}.$$

Прологарифмируем:

$$\ln\mathcal{L}(x_1,\dots,x_m,p,q,n)=\sum_{t=1}^{m}\Bigg(-\ln x_t!+(n-x_t)\ln(1-p)+x_tq+$$
$$+\ln\left(\sum_{j=1}^{x_t}\frac{n!}{(n-j)!}c(x_t,j)(p\alpha)^j(1-p)^{(x_t-j)}\right)\Bigg).$$

Применим метод моментов для биномиально-логарифмического распределения, то есть если нам известно рассеяние, то можем выразить p через q:

$$eS_N=\frac{\ln(1-q)+qp}{(1-q)\ln(1-q)}\implies p=\frac{\ln(1-q)(1-q)\cdot eS_N-\ln(1-q)}{q}.$$

А теперь для логарифмически-биномиального распределения, но для него через математическое ожидание, так как через рассеяние это не представляется возможным:

$$p=-\frac{\ln(1-q)(1-q)}{nq}\cdot \mathsf{E}S_N.$$

Поэтому, подставив эти выражения в функцию правдоподобия, получим функцию от двух, а не трёх параметров:

$$\ln \mathcal{L}(x_1, \ldots, x_m, q, n).$$

Для нахождения оценки максимального правдоподобия необходимо продифференцировать это выражение и найти нуль производной, однако для данного случая будем оценивать параметры q и n численными методами многомерной оптимизации (функция `optim` в R).

Например, in vitro при 35 Гр получаем наилучшие оценки: $n = 78.693, q = 0.188, p = 0.016$. Значимость согласия по критерию χ^2: $p - value = 0.26$. Результат представлен на рис. 2.2.

2.4. Применение в радиобиологии и интерпретация

Оценки параметров и значимости критерия хи-квадрат по данным in vivo и in vitro.

Таблица 2.1. In vivo, логарифмически-биномиальное распределение

Гр	n	q	p	**p-v**
0	5.00	0.20	0.07	**0.18**
5	5.00	0.20	0.12	**0.73**
10	5.00	0.20	0.15	**0.82**
15	5.00	0.20	0.21	**0.03**
20	5.00	0.20	0.30	**0.01**
25	5.00	0.20	0.19	**0.39**
30	5.79	0.53	0.21	**0.11**
35	5.79	0.43	0.25	**0.07**
40	5.62	0.47	0.28	**0.06**
45	5.52	0.46	0.32	**0.01**

Таблица 2.2. In vitro, биномиально-логарифмическое распределение

Гр	n	q	p	**p-v**
0	12.57	0.00	0.03	**0.59**
5	6.76	0.25	0.04	**0.87**
10	21.02	0.08	0.03	**0.07**
15	15.72	0.38	0.04	**0.20**
20	105.3	0.20	0.00	**0.33**
25	64.32	0.17	0.02	**0.27**
30	151.6	0.27	0.01	**0.32**
35	78.69	0.19	0.02	**0.26**
40	81.41	0.10	0.02	**0.25**

Логарифмически-биномиальное распределение мы применили к данным in vivo, а биномиально-логарифмическое — in vitro. С помощью численной оптимизации двухмерных функций правдоподобия для каждой выборки ядерных аномалий у крыс получены результаты согласованности по критерию χ^2, отражённые в таблице 2.1.

Можно предположить следующую интерпретацию параметров: n — экстенсивность внешнего воздействия (экстенсивность облучения), p — интенсивность внешнего воздействия (вероятность возникновения аномалии при облучении), q — инертность (вероятность развития аномалии при делении). Таким образом, наследование аномалий осуществляется по логарифмическому закону, а образование аномалий за счет облучения по биномиальному.

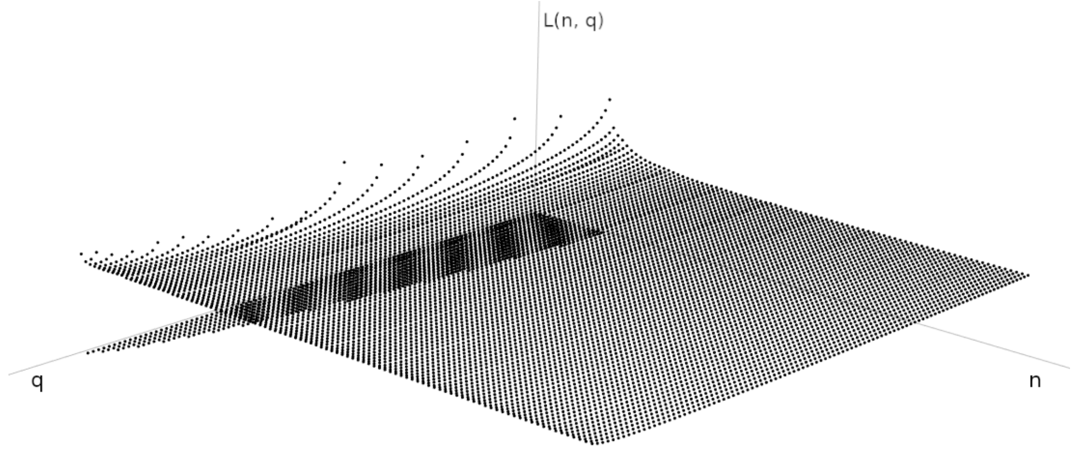

Рис. 2.3. Функция правдоподобия для логарифмически-биномиального распределения в случае 20 Гц in vivo

Однако если внимательнее посмотреть на результаты in vivo, то увидим почти равные значения параметра n (примерно начальное значение при оптимизации). И это не удивительно, так как, график функции правдоподобия логарифмически-биномиального распределения для всех выборок, имеет малую изменчивость (рис. 2.3): минимум достигается во многих точках, поэтому почти при любом начальном значении n и q будет получено одинаковое согласие. Это может говорить об излишнем количестве параметров или вообще о неподходящем виде распределения для эмпирических данных.

При этом согласованность биномиально-логарифмического распределения с данными in vivo по p-value по большей части выборок больше (таблица 2.3), чем у логарифмически-биномиального, поэтому можно предположить о том, что порядок облучения и образования ядерных аномалий не так важен, как само их наличие в эксперименте с нужным периодом образования и мощностью излучения.

Таблица 2.3. Оценки параметров и значимости критерия хи-квадрат по данным in vivo, биномиально-логарифмическое распределение

Гр	n	q	p	**p-v**
0	3.25	0.00	0.12	**0.13**
5	6.52	0.00	0.10	**0.81**
10	6.33	0.00	0.13	**0.21**
15	5.45	0.00	0.21	**0.01**
20	96.35	0.00	0.02	**0.03**
25	123.1	0.00	0.01	**0.55**
30	54.91	0.16	0.03	**0.27**
35	136.55	0.04	0.01	**0.03**
40	135.52	0.02	0.02	**0.10**
45	43.30	0.00	0.06	**0.05**

2.5. Применение к анализу встречаемости слов

2.5.1. Постановка задачи

Для грамотного построения речевых нейросетей немаловажным является знание о распределении встречаемости слов в тексте. Хотелось бы найти такую модель (или набор моделей), которая отвечала бы всевозможным типам слов, какими бы они ни были. Сначала формализуем данную задачу.

Дан текст из n глав. Некоторое слово ω_j встречается в i-ой главе $x_i^{(j)}$ раз. Вопрос: какому распределению удовлетворяет выборка $(x_1^{(j)}, \ldots, x_n^{(j)})$?

В работе [2] утверждается, что неплохое согласование даёт модель отрицательного бинома. Однако есть ряд слов, не согласующихся с отрицательно-биномиальным распределением, поэтому возникает идея проверить их согласованность с биномиально-логарифмическим, как обобщением отрицательно-биномиального. По крайней мере, если отрицательно-биномиальное распределение даёт хорошую согласованность, то и биномиально-логарифмическое должно давать похожий результат.

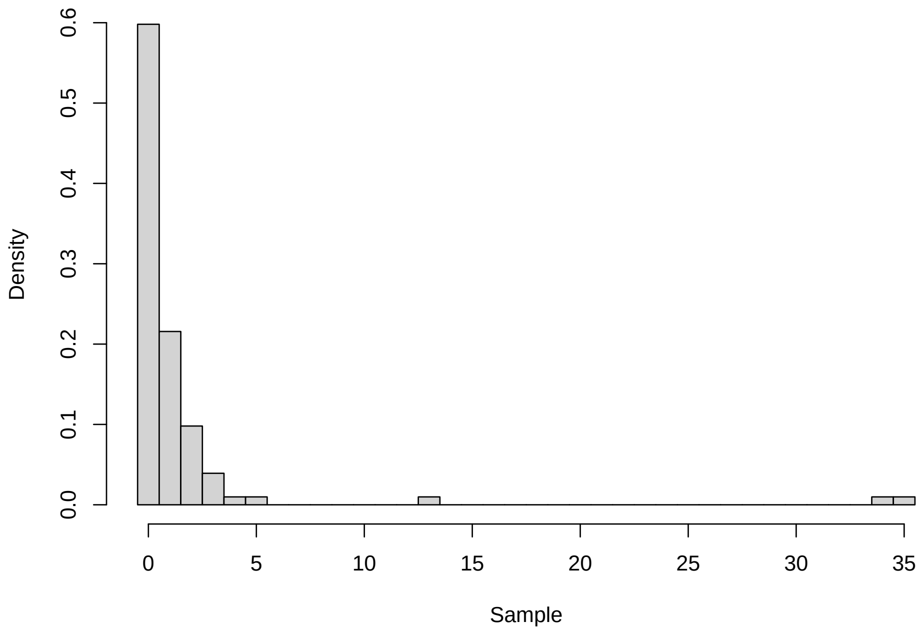

Рис. 2.4. Гистограмма распределения слова "coat"

2.5.2. Применение биномиально-логарифмическое к анализу текста

В качестве исследуемого объекта был взят роман американского писателя Теодора Драйзера „Американская трагедия“ на английском языке. С помощью самописной программы получены выборки частот встречаемости слов по главам, которых в данном произведении 102 (n). Всего оказалось 14134 ($\omega_j,\ j \in \overline{1:14134}$) различных слов, однако большая их часть встречается не более 5–6 раз во всём тексте. Поэтому будем рассматривать первые (по общему количеству во всём тексте) 1000 слов, а также не будем касаться слов, встречаемость которых в одной главе превосходит 50, так как это сопряжено с проблемами вычисления вероятностей биномиально-логарифмического распределения.

Насчёт вычисления вероятностей: каждая из них является суммой произведений некоторого факториала, делённого на меньший факториал, чисел Стирлинга и чисел меньше единицы в степени, зависящей от индекса в сумме. Получается, что мы оперируем крайне быстро растущими и убывающими функциями, что сказывается на погрешности вычисления. Например, число Стирлинга 50-ой степени имеет 63-ий порядок. Всё это не даёт использовать в практических вычислениях биномиально-логарифмическое распределение для всевозможных эмпирических распределений без какой-то существенной модификации.

Согласованность для отрицательно-биномиальное распределение проверялась по критерию χ^2 с оцениванием параметров по методу максимального правдоподобия. Согласованность для биномиально-логарифмическое проверялась аналогично случаю в радиобиологии.

По итогу, количество слов, для которых критерий для обоих распределений давал p-value больше 0.2, равняется 605, при этом отдельно для отрицательно-биномиальное распределение — 666, а для биномиально-логарифмическое — 607, то есть разница небольшая. Значит, можем предположить верность утверждения выше: согласованность для отрицательно-биномиального распределения равносильна согласованности для биномиально-логарифмического.

Основная часть, для которой не было достигнуто согласование — это слова, распределение которых имеет „тяжёлый“ правый хвост, выражающийся в элементах выборки, которые имеют значение сильно больше остальных. Пример такого слова представлен на рис. 2.4.

Заключение

Многие случайные процессы подчиняются некоторым известным и простым распределениям, однако, иногда для их описания требуется усложнять модели, вводя, например, понятие смеси распределений или, как в нашем случае, суперпозицию более простых. К сожалению, они наследуют некоторые свойства своих образующих, как это было показано с рассеянием в лемме 1. Поэтому особый интерес представляют такие комбинации распределений, которые дают широкое разнообразие своих характеристик и форм.

В качестве такого распределения мною было рассмотрено биноминально-логарифмическое и логарифмически-биномиальное распределения, для которых были найдены математическое ожидание, дисперсия, рассеяние (и при каких параметрах его логарифм меняет знак), вероятности до $k = 4$ для логарифмически-биномиального и общая формула вероятностей для биномиально-логарифмического, а также показана применимость на радиобиологических данных из статьи [1]. В работе для нахождения оптимальных параметров, дающих наибольшее согласование были применены метод моментов и метод максимального правдоподобия реализуемый численными методами многомерной оптимизации. Также была предложена интерпретация параметров для биномиальной и логарифмической компонент, как экстенсивность и интенсивность внешнего воздействия и инертность, соответственно.

Была предпринята попытка применения биномиально-логарифмическое к задаче встречаемости слов в тексте. Однако пока удалось только показать некоторую равнозначность применения отрицательно-биномиального распределения и биномиально-логарифмического для данной задачи. Они оба дают плохую согласованность для слов, которые имеют значения частот сильно больше математического ожидания или, по-другому, „тяжёлые“ хвосты.

Список литературы

1. Динамика роста числа ядерных аномалий рабдомиосаркомы RA-23 при увеличении дозы острого редкоионизирующего облучения. Исследование на основе модели реинтрантно-биномиального распределения / Алексеева Н. П., Алекссев А. О., Вахтин Ю. Б., Кравцов В. Ю., Кузоватов С. Н. и Скорикова Т. И. // Цитология. — 2008. — С. 528–534.
2. Alexeeva N. Sotov A. The Negative Binomial Model of Word Usage // Electronic Journal of Applied Statistical Analysis. — 2013. — Vol. 6, no. 1.
3. Феллер В. Введение в теорию вероятностей и её приложения. В 2 т. — Москва : Мир, 1952. — Т. 1.
4. Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики. — Москва : Мир, 1998. — ISBN: 5-03-001793-3.