Санкт-Петербургский государственный университет
Прикладная математика и информатика

Учебная практика 3 (научно-исследовательская работа)(семестр 6)

«Оценка параметров сложных распределений с применением в радиобиологии»

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 A

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.

Моя задача состоит в проверке согласованности эмпирического распределения, полученного в работе Алексеевой [1], с различными сложными распределениями. Необходимо рассчитать их параметры, промоделировать, вычислить критерий согласованности с эмпирическим распределением и сравнить с моделью реинтрантно-биномиального распределения.

Глава 1

Анализ сложных распределений

1.1. Реинтрантно-биномиальное распределение

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 [2]. 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 будет

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

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 биномиальных распределений.

Также в статье представлены формулы вычисления математического ожидания и дисперсии:

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

Оно хорошо описывало экспериментальные данные, однако имело целых четыре параметра, которые при интерпретации сводились к двум произведениям реинтрантных компонент, что говорило о возможном упрощении его структуры.

1.2. Производящие функции Пуассоновских распределений

Первым делом возникла идея о замене реинтрантно-биномиального закона на Пуассоновский. Они имеют много замечательных свойств и хорошо моделируются.

Сначала введём производящую функцию Пуассоновского распределения:

$$f(t) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{(\lambda t)^k}{k!} = e^{-\lambda + \lambda t}.$$

В первую очередь мною будет рассматриваться случай двойного Пуассоновского распределения, чья производящая функция имеет следующий вид:

$$h(t) = e^{-\lambda_1 + \lambda_1 e^{-\lambda_2 + \lambda_2 t}} = \sum_{k=0}^{\infty} e^{-\lambda_1} \frac{\left(\lambda_1 \sum\limits_{n=0}^{\infty} e^{-\lambda_2} \frac{(\lambda_2 t)^n}{n!}\right)^k}{k!}.$$

Математическое ожидание двойного Пуассоновского распределения (производная производящей функции в единице) по свойству производящих функций [2, с. 272]:

$$\mathsf{E}\xi = h'(1) = \left(e^{-\lambda_1 + \lambda_1 e^{-\lambda_2 + \lambda_2 t}}\right)'(1) = \left(\lambda_1 e^{-\lambda_1 + \lambda_1 e^{-\lambda_2 + \lambda_2 t}} \cdot \lambda_2 e^{-\lambda_2 + \lambda_2 t}\right)(1) = \lambda_1 \lambda_2.$$

Вторая производная производящей функции двойного Пуассоновского распределения:

$$h''(t) = \left(e^{-\lambda_1 + \lambda_1 e^{-\lambda_2 + \lambda_2 t}}\right)'' = \lambda_1 \lambda_2^2 e^{-\lambda_1 + \lambda_1 e^{-\lambda_2 + \lambda_2 t}} \cdot e^{-\lambda_2 + \lambda_2 t} \left(\lambda_1 e^{-\lambda_2 + \lambda_2 t} + 1\right),$$

тогда дисперсия по свойствам производящих функций равна:

$$\mathsf{D}\xi = h''(1) + h'(1) - (h'(1))^2 = \lambda_1 \lambda_2^2 (\lambda_1 + 1) + \lambda_1 \lambda_2 - (\lambda_1 \lambda_2)^2 = \lambda_1 \lambda_2 (\lambda_2 + 1).$$

1.2.1. Моделирование двойного Пуассоновского распределения

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

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

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 $P\{X_j = x\}$ [2, с. 291].

Чтобы смоделировать выборку из n элементов двойного Пуассоновского распределения, нужно смоделировать выборку из n элементов первого Пуассоновского распределения, для каждого из которых вычисляется сумма выборки из элементов второго, чьё количество равно значению элемента из первого.

В двойном Пуассоне: $N \in \pi(\lambda_1)$ и $X_j \in \pi(\lambda_2)$. Функция для моделирования двойного Пуассоновского распределения в R:

```
rppois <- function(n = 1, lambda1 = 1, lambda2 = 1) {
    res <- c()

    for(i in rpois(n, lambda1))
        res <- c(res, sum(rpois(i, lambda2)))

    return(res)
}
```

Результат промоделированного двойного распределения Пуассона с $\lambda_1 = 3$ и $\lambda_2 = 2$ на рис. 1.1. Значения математического ожидания и дисперсии согласуются с теорией:

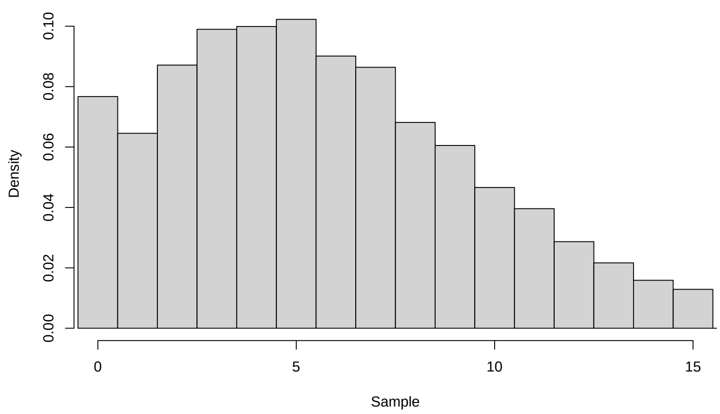

Рис. 1.1. Промоделированное двойное распределение Пуассона с $\lambda_1 = 3$ и $\lambda_2 = 2$

$$\mathsf{E}\xi_n = 6.0004 \approx 6 = \mathsf{E}\xi,$$

$$\mathsf{D}\xi_n = 18.13981 \approx 18 = \mathsf{D}\xi.$$

1.2.2. Вероятности двойного Пуассоновского распределения

Для проверки согласованности двух распределений необходимо знать их вероятности. Для этого вычислим их через преобразование производящей функции:

$$\sum_{n=0}^{\infty} e^{-\lambda_1} \frac{\left(\lambda_1 \sum\limits_{k=0}^{\infty} e^{-\lambda_2} \frac{(\lambda_2 t)^k}{k!}\right)^n}{n!} = \sum_{n=0}^{\infty} \frac{e^{-\lambda_1}\left(\lambda_1 e^{-\lambda_2}\right)^n}{n!} \cdot e^{n\lambda_2 t} = \sum_{n=0}^{\infty} \frac{e^{-\lambda_1}\left(\lambda_1 e^{-\lambda_2}\right)^n}{n!} \cdot$$
$$\cdot \sum_{k=0}^{\infty} \frac{(\lambda_2 n t)^k}{k!} = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} \frac{e^{-\lambda_1}\left(\lambda_1 e^{-\lambda_2}\right)^n}{n!} \cdot \frac{(\lambda_2 n t)^k}{k!} =$$
$$= \sum_{k=0}^{\infty} \frac{e^{-\lambda_1} t^k \lambda_2^k}{k!} \cdot \sum_{n=0}^{\infty} \frac{n^k \left(\lambda_1 e^{-\lambda_2}\right)^n}{n!}.$$

Таким образом, получим для $k = 0, 1, 2, \ldots$ такие вероятности:

$$P(X = k) = \frac{e^{-\lambda_1}\lambda_2^k}{k!} \cdot \sum_{n=0}^{\infty} \frac{n^k \left(\lambda_1 e^{-\lambda_2}\right)^n}{n!},$$

что можно использовать в численных методах с некоторой погрешностью.

Другой способ получения вероятностей — это получение рекурсивной формулы для вероятностей, опираясь на тот факт, что если $h(t)$ — производящая функция, то вероятности распределения этой функции будет вычисляться так:

$$P(X = j) = \frac{h^{(j)}(0)}{j!}.$$

Для первых двух производных результат уже известен:

$$P(X = 1) = h'(0) = \lambda_1\lambda_2 e^{-\lambda_1 + \lambda_1 e^{-\lambda_2}} \cdot e^{-\lambda_2};$$
$$P(X = 2) = \frac{h''(0)}{2} = \frac{1}{2}\lambda_1\lambda_2^2 e^{-\lambda_1 + \lambda_1 e^{-\lambda_2}} \cdot e^{-\lambda_2}\left(\lambda_1 e^{-\lambda_2} + 1\right).$$

Это база математической индукции, далее определим по индукции:

$$P(X = n) = \frac{h^{(n)}(0)}{n!} = \frac{1}{n!}\lambda_1\lambda_2^n e_{\lambda_1} \cdot e_{\lambda_2} \sum_{k=0}^{n-1} a_{kn} e_{\lambda_2}^k \lambda_1^k,$$

$$a_{0n} = 1, a_{(n-1)n} = 1, a_{kn} = (k+1)a_{k(n-1)} + a_{(k-1)(n-1)}, k = 1, 2, \ldots, n-2,$$

где $e_{\lambda_1} = e^{-\lambda_1 + \lambda_1 e^{-\lambda_2}}$, $e_{\lambda_2} = e^{-\lambda_2}$. Проверяется непосредственным дифференцированием.

1.3. Тройное Пуассоновское распределение

Однако более интересным объектом для исследования представляется тройное Пуассоновское распределение, имеющее производящую функцию вида:

$$g(t) = e^{-\lambda_1+\lambda_1 e^{-\lambda_2+\lambda_2 e^{-\lambda_3+\lambda_3 t}}},$$

или суперпозиция трёх производящих функций Пуассоновского распределения.

Для удобства записи обозначим:

$$e_{\lambda_1}(t) = e^{-\lambda_1+\lambda_1 e^{-\lambda_2+\lambda_2 e^{-\lambda_3+\lambda_3 t}}};$$
$$e_{\lambda_2}(t) = e^{-\lambda_2+\lambda_2 e^{-\lambda_3+\lambda_3 t}};$$
$$e_{\lambda_3}(t) = e^{-\lambda_3+\lambda_3 t};$$

и получим математическое ожидание для данного распределения:

$$\mathsf{E}\nu = h'(1) = (e_{\lambda_1})'(1) = (\lambda_1 e_{\lambda_1} \cdot \lambda_2 e_{\lambda_2} \cdot \lambda_3 e_{\lambda_3})(1) = \lambda_1\lambda_2\lambda_3.$$

Теперь посчитаем вторую производную:

$$h''(t) = \lambda_1\lambda_2\lambda_3^2 e_{\lambda_1} e_{\lambda_2} e_{\lambda_3}(\lambda_1\lambda_2 e_{\lambda_2} e_{\lambda_3} + \lambda_2 e_{\lambda_3} + 1).$$

Тогда дисперсия распределения равна:

$$\begin{aligned}\mathsf{D}\nu =& h''(1) + h'(1) - (h'(1))^2 = \lambda_1\lambda_2\lambda_3^2(\lambda_1\lambda_2 + \lambda_2 + 1) + \lambda_1\lambda_2\lambda_3 - (\lambda_1\lambda_2\lambda_3)^2 = \\ =& \lambda_1\lambda_2\lambda_3(\lambda_2\lambda_3 + \lambda_3 + 1).\end{aligned}$$

Программа для моделирования тройного Пуассоновского распределения в R:

```
rpppois <- function(n = 1, lambda1 = 1, lambda2 = 1, lambda3 = 1){
    res <- c()

    for(i in rpois(n, lambda1))
    {
        summa <- 0

        for(j in rpois(i, lambda2))
            summa <- sum(summa, sum(rpois(j, lambda3)))
```

```
        res <- c(res, summa)
    }

    return(res)
}
```

1.3.1. Рассеяние пуассоновских распределений

Несмотря на то что рассмотренные выше распределения вызывают интерес, как бесконечно делящиеся распределения, они всё же не подошли нам в качестве описательных моделей по причине перерассеяности. Для двойного:

$$e\xi = \frac{\mathsf{D}\xi}{\mathsf{E}\xi} = \lambda_2 + 1,$$

для тройного:

$$e\nu = \lambda_2\lambda_3 + \lambda_3 + 1,$$

а ни при каких $\lambda_1, \lambda_2, \lambda_3 > 0$ оно не становится меньше 1.

1.4. Логарифмическое распределение

Единственное (из известных) распределений имеющих без составления в сложные распределения переменный знак логарифма рассеяния — это логарифмическое распределение:

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

Оно имеет производящую функцию:

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

Вычислим математическое ожидание:

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

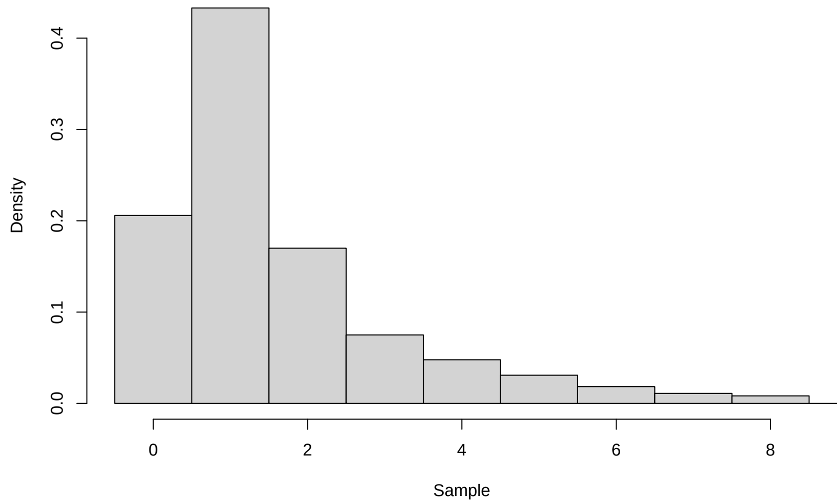

Рис. 1.2. Промоделированное логарифмическое распределение с параметром. $p_0 = 0.2$ и $p = 0.75$

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 может менять знак.

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

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

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:

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

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. 1.2.

Если мы посмотрим на данные в статье [1], то убедимся, что максимальная частота всегда на значениях 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.

1.5. Рассеяние случайной величины

В статье Алексеевой [1] показано, что эмпирические данные должны иметь модель распределения, рассеяние которой имеет логарифм переменного знака, 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.

Тогда хотелось бы понять, как ведёт себя рассеяние случайной величины, удовлетворяющей сложному закону распределения.

Утверждение 1. *Пусть S_N — сумма случайного числа N случайных величин ξ_i, одинаково распределённых. 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:

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

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 утверждения 1 следует, что, так как мы рассматриваем дискретные распределения с неотрицательными значениями случайной величины, то при рассеянии 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, внутренне 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

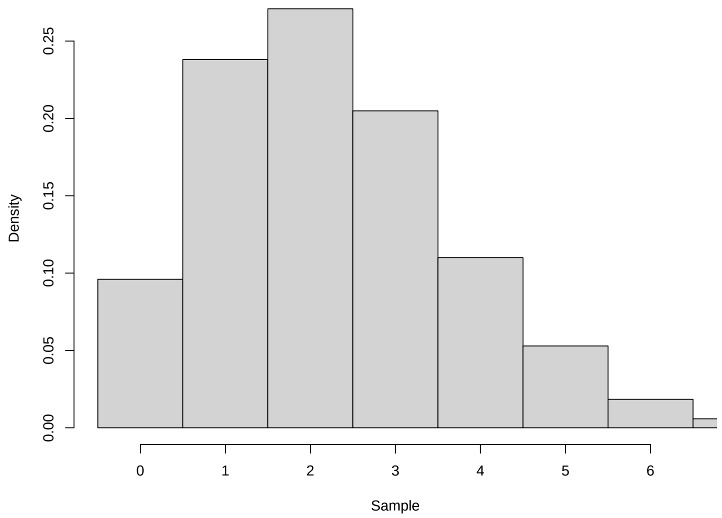

Рис. 2.1. Промоделированное биномиально-логарифмическое распределение. $q = 0.2, p = 0.25$ и $n = 8$

В качестве такого распределения рассмотрим суперпозицию биномиального и логарифмического распределений, производящая функция которого имеет вид:

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

Вычислим основные его характеристики через характеристики образующих его распределений:

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

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

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

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

2.1.1. Вероятности

Как известно из свойств производящей функции: 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:

$$P(S_N=k)=\left(\frac{d^n}{dx^n}h(g)\right)(0)\cdot\frac{1}{n!}=\sum\frac{1}{m_1!m_2!\cdots m_n!}h^{(m_1+m_2+\cdots+m_n)}(0)\cdot$$
$$\cdot\prod_{j=1}^{n}\left(\frac{g^{(j)}(0)}{j!}\right)^{m_j}=$$
$$=\sum\frac{1}{m_1!m_2!\cdots m_n!}(m_1+m_2+\cdots+m_n)!\cdot P(N=m_1+m_2+\cdots+m_n)\cdot$$
$$\cdot\prod_{j=1}^{n}\big(P(\xi=j)\big)^{m_j},$$

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

$$0!P(S_N=0)=P(N=0)$$
$$1!P(S_N=1)=P(N=1)\cdot P(\xi=1)$$
$$2!P(S_N=2)=2!P(N=2)\cdot(P(\xi=1))^2+P(N=1)\cdot 2!P(\xi=2)$$
$$3!P(S_N=3)=3!P(N=3)\cdot(P(\xi=1))^3+3\cdot 2!P(N=2)\cdot P(\xi=1)\cdot 2!P(\xi=2)+$$
$$+P(N=1)\cdot 3!P(\xi=3)$$
$$4!P(S_N=4)=4!P(N=4)\cdot(P(\xi=1))^4+6\cdot 3!P(N=3)\cdot(P(\xi=1))^2\cdot 2!P(\xi=2)+$$
$$+3\cdot 2!P(N=2)\cdot(2!P(\xi=2))^2+4\cdot 2!P(N=2)\cdot P(\xi=1)\cdot 3!P(\xi=3)+$$
$$+P(N=1)\cdot 4!P(\xi=4).$$

Преобразуем, подставив вероятности соответствующих распределений:

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

Тогда общая формула принимает вид:

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

где $\alpha = \frac{-1}{\ln(1-q)}$, $c(k,j)$ — число Стирлинга первого рода без знака [3], задающее число перестановок из k элементов с j циклами. Но нам будет достаточно первых пяти вероятностей для оценок распределения.

2.2. Метод максимального правдоподобия

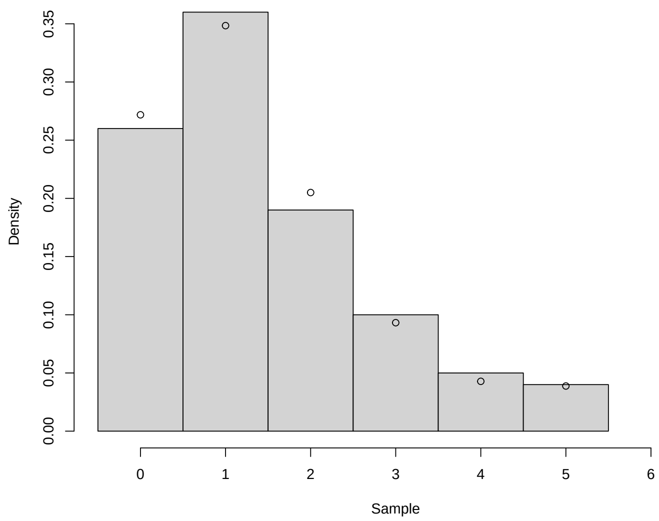

Рис. 2.2. Эмпирические частоты (столбики) и теоретические частоты (точки), вычисленные по методу максимального правдоподобия

Параметры распределения будем находить по методу максимального правдоподобия. Составим функцию правдоподобия:

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

Прологарифмируем:

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

Применим метод моментов, то есть если нам известно математическое ожидание, то можем выразить p через q:

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

Поэтому, подставив это выражение 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 параметр q численными методами, а n переберём или предположим модель.То есть при фиксированном n будем находить максимум у функции:

$$\ln \mathcal{L}(x_1, \ldots, x_m, q, n) = \sum_{t=1}^{m} \ln \big(P(S_N = x_t, p) \big).$$

Например, in vitro при 35 Гр получаем наилучшие оценки при $n = 2$: $q = 0.536, p = 0.479$. Значимость согласия по критерию χ^2: $p - value = 0.933$. Результат представлен на рис. 2.2.

2.3. Применение в радиобиологии и интерпретация

Биномиально-логарифмическое распределение мы применили к данным in vitro. Предположительно вместо перебора параметра n была выбрана модель почти экспоненциального $\left(\frac{e^{t-1}}{t}\right)$ роста, для которой 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 (вероят-

Таблица 2.1. Оценки параметров и значимости критерия хи-квадрат по данным in vitro.

	Доза, Гр	n	q	P	p.value
1	0	1.00	0.32	0.32	0.46
2	5	1.36	0.39	0.20	0.74
3	10	2.46	0.25	0.21	0.05
4	15	5.02	0.44	0.13	0.21
5	20	10.92	0.22	0.05	0.38
6	25	24.74	0.22	0.04	0.21
7	30	57.63	0.26	0.02	0.27
8	35	137.08	0.24	0.01	0.54
9	40	331.22	0.16	0.00	0.53

ность развития аномалии при делении). Таким образом, наследование аномалий осуществляется по логарифмическому закону, а образование аномалий за счет облучения по биномиальному.

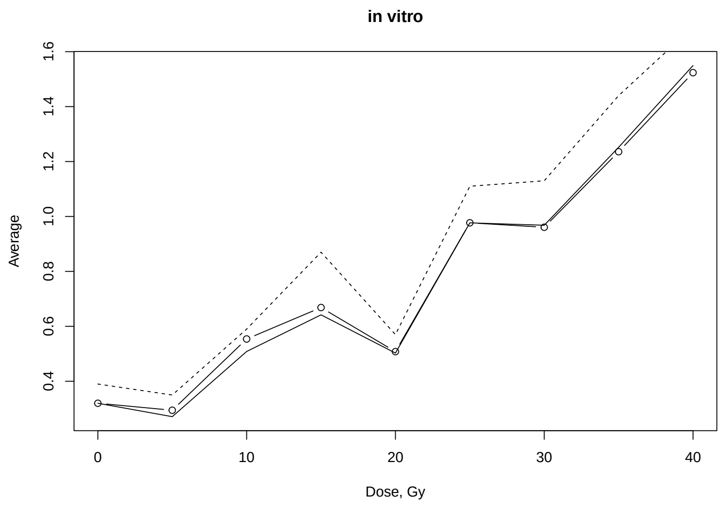

Рис. 2.3. Сравнение среднего и $n \cdot p$ для разных моделей (сплошная линия — почти экспоненциальная, линия с точками — линейная, пунктирная — среднее).

Также была рассмотрена модель линейного роста параметра n, однако как видно из рис. 2.3 это почти не влияет на значение $n \cdot p$, которое, в свою очередь, несильно отличается от среднего количества аномалий, что говорит о небольшом влиянии параметра q, как показателе образования аномалий при делении.

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 утверждении 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, для которого были найдены математическое ожидание, дисперсия, рассеяние (и при каких параметрах его логарифм меняет знак), общая формула вероятностей, 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. Также была предложена интерпретация параметров для биномиальной и логарифмической компонент, как экстенсивность и интенсивность внешнего воздействия и инертность, соответственно.

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. Феллер В. Введение в теорию вероятностей и её приложения. В 2 т. — Москва : Мир, 1952. — Т. 1.
3. Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики. — Москва : Мир, 1998. — ISBN: 5-03-001793-3.