Санкт-Петербургский государственный университет

**Олейник Михаил Владимирович**

**Выпускная квалификационная работа**

# Оценка параметров сложных распределений с применением в радиобиологии и лингвистике

Уровень образования: бакалавриат

Направление 01.03.02 «Прикладная математика и информатика»

Основная образовательная программа СВ.5004.2018 «Прикладная математика и информатика»

Научный руководитель:

Доцент, кафедра статистического моделирования

кандидат ф.-м. н., доцент Н. П. Алексеева

Рецензент:

Научный сотрудник, СПбГУ, факультет математики и компьютерных наук

Ю. С. Белоусов

Санкт-Петербург

2024

Saint Petersburg State University
Applied Mathematics and Computer Science

**Oleynik Michael Vladimirovich**

**Graduation Project**

# Estimation of parameters of complex distributions with applications in radiobiology and linguistics

Scientific Supervisor:
Docent, Department of Statistical Modelling
N. P. Alexeeva

Reviewer:
Researcher, St. Petersburg State University,
Faculty of Mathematics and Computer
Science Y. S. Belousov

Saint Petersburg
2024

# Оглавление

# Введение

В статистическом анализе при подборе модели для эмпирических данных важным критерием является её сложность с точки зрения количества параметров, скорости их оценки и вычислительных затрат. Сведение задачи к уже известным простым распределениям представляет идеальную ситуацию, однако, ввиду комплексности многих реальных задач, верной может оказаться только достаточно сложная модель с несколькими параметрами, оценка которых представляет собой отдельную область изучения. Поэтому при достижении достаточного согласования с некоторой сложной моделью следует попытаться взять, например, частный случай этой модели и проверить её или, что бывает достаточно часто, взять совершенно другую, но немного проще по какому-либо параметру сравнения.

В данной работе были взяты за основу модели сложных распределений (случайная сумма случайных величин), которые нашли широкое применение в описании ветвящихся процессов [1]. Они хорошо исследованы и имеют широкое применение в физике частиц, но многие из них ограничены в применении из-за перерассеяность (рассеяние больше 1), что не позволяет использовать их в ситуациях, когда логарифм рассеяния имеет переменный знак.

Другое применение сложных распределений часто можно увидеть в комплексных биологических системах [2], где каждая из составляющих отвечает за более простую часть общего процесса.

Наконец, популярными моделями страхования имущества являются модели Крамера-Лундеберга с непрерывными или стохастическими премиями [3], в которых учёт несчастных случаев и соответственно выплат производится в качестве случайной суммы случайных величин, где количество выплат распределено по одному закону, а размер выплат — по другому.

В работе Алексеевой [4] 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 vivo и in vitro через 52 часа после X-облучения в дозах 5–45 Гр.

Анализ текстов [5] показал применимость модели отрицательно-биномиального распределения к встречаемости по главам многих слов. Однако некоторые из них дают плохую согласованность из-за специфического вида эмпирических распределений.

Цель выпускной квалификационной работы состоит в проверке согласованности эмпирических распределений, полученных в работе Алексеевой [4], с различными сложными распределениями, которые также следует применить в анализе встречаемости слов в тексте. Необходимо вычислить их вероятности, оценить параметры этих распределений, промоделировать, вычислить критерий согласованности с эмпирическими данными и сравнить с моделью реинтрантно-биномиального распределения для случая радиобиологии и моделью отрицательного бинома для случая анализа текстов.

Актуальность работы обусловлена широкими возможностями применением сложных статистических моделей в современном мире больших данных.

Для достижения обозначенных целей в работе были использованы различные статистические методы: для оценки параметров — метод моментов и максимального правдоподобия, для проверки согласия с эмпирическими распределениями — критерий $\chi^2$, для доказательства формул вероятностей сложных распределений — свойства производящих функций и метод математической индукции.

Работа состоит из введения, трёх основных глав, заключения и приложения с кодом `R`. В первой главе вводятся основные понятия, исследуется не самое популярное логарифмическое распределение и поясняется метод отбора сложных распределений. Во второй главе исследуются три составных распределения: биномиально-логарифмическое, логарифмически-биномиальное и частный случай второго — логарифмически-пуассоновское. Вычисляются их вероятности и исследуется согласование с радиобиологическими данными. В третьей главе рассматривается задача встречаемости слов в тексте с применением биномиально-логарифмического распределения.

# 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 [6]. Однако, если коэффициентами ряда будут вероятности некоторого дискретного распределения:

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

то

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

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 распределений, а также для вычисления вероятностей нам понадобятся свойства производящей функции [6] ($\xi$ — 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. Рассеяние случайной величины

Помимо математического ожидания и дисперсии для характеризации случайной величины может использоваться другая величина на их основе — рассеяние.

**Определение 1.** *Пусть $\xi$ — случайная величина с математическим ожиданием $\mathsf{E}\xi$ и дисперсией $\mathsf{D}\xi$. Тогда рассеянием называется отношение дисперсии к математическому ожиданию:*

$$e(\xi) = \frac{\mathsf{D}\xi}{\mathsf{E}\xi}.$$

Смысл этой величины состоит в том, насколько при изменении параметров распределения меняется величина дисперсии по отношению к математическому ожиданию.

Значение этой величины может быть крайне важно, так как многие известные распределения имеют величину логарифма рассеяния (в дальнейшем нам будет удобно оперировать именно ей) либо больше, либо меньше нуля (то есть либо дисперсия, либо математическое ожидание растёт быстрее), а эмпирические данные могут иметь совершенно произвольную величину рассеяния.

Например, для распределения Пуассона ($\xi \sim Pois(\lambda), P(\xi = k) = \frac{\lambda^k}{k!}e^{-\lambda}$) ситуация самая простая:

$$\ln e(\xi) = \ln \frac{\mathsf{D}\xi}{\mathsf{E}\xi} = \ln \frac{\lambda}{\lambda} = 0.$$

Для биномиального распределения ($\xi \sim Bin(n, p), P(\xi = k) = C_n^k p^k (1-p)^{n-k}$):

$$\ln e(\xi) = \ln \frac{np(1-p)}{np} = \ln(1-p), \quad 0 \leqslant p \leqslant 1,$$

то есть меньше 0.

Поэтому распределения с логарифмом рассеяния меньше или больше нуля могут быть сразу исключены из потенциальных моделей для конкретных эмпирических данных.

### 1.1.3. Сложные распределения

**Определение 2.** *Под сложным распределением будем понимать конструкцию следующего вида:*

$$S_\tau = \xi_1 + \cdots + \xi_\tau,$$

*где* $\xi_i$ $\forall i \geqslant 0$ *— случайные величины независимые и одинаково распределённые,* $\tau$ *— случайная величина независимая от* $\xi_i$*.*

Если $\xi_i$ имеют производящую функцию $g(t)$, а $\tau$ — $f(t)$, то производящей функцией сложного распределения на их основе, то есть $S_\tau$, является $f(g(t))$ [6].

### 1.1.4. Реинтрантно-биномиальное распределение

Реинтрантно-биномиальное распределение, используемое в работе [4], имеет производящую функцию вида:

$$h_0(t) = (p_0(p_1 t + q_1)^{n_1} + q_0)^{n_0},$$

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 — простое, то есть не являющееся суперпозицией или смесью, дискретное распределение, имеющее без составления в сложные

распределения переменный знак логарифма рассеяния — это логарифмическое распределение ($\xi \sim Log(q)$) с такими вероятностями:

$$P(\xi = k) = \frac{\alpha q^k}{k}, \quad \alpha = \frac{-1}{\ln(1-q)}, q \in (0,1), k = 1, 2, \ldots$$

Оно имеет производящую функцию:

$$g_1(t) = -\alpha \ln(1 - qt).$$

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

$$\mathsf{E}\xi = g_1'(1) = \left(\frac{\alpha q}{1-qt}\right)(1) = \frac{\alpha q}{1-q},$$

и дисперсию:

$$g_1''(t) = \frac{\alpha q^2}{(1-qt)^2}$$

$$\mathsf{D}\xi = g_1''(1) + g_1'(1) - (g_1'(1))^2 = \frac{\alpha q^2}{(1-q)^2} + \frac{\alpha q}{1-q} - \frac{\alpha^2 q^2}{(1-q)^2} = \alpha q \cdot \frac{1-\alpha q}{(1-q)^2}.$$

Тогда рассеяние равно:

$$e\xi = \frac{\mathsf{D}\xi}{\mathsf{E}\xi} = \frac{1-\alpha q}{1-q}.$$

Найдём при каком $q$ логарифм рассеяния меняет знак:

$$e\xi = \frac{\mathsf{D}\xi}{\mathsf{E}\xi} = \frac{1-\alpha q}{1-q} = 1$$

$$1 - \alpha q = 1 - q$$

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

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

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

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

То есть логарифм рассеяния при $q \in (0,1)$ 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 $q_0$:

$$P(\xi = 0) = q_0;$$

$$P(\xi = k) = (1 - q_0) \cdot \frac{\alpha q^k}{k}, \quad q \in (0,1), k = 1, 2, \ldots.$$

Производящая функция этого распределения:

$$g_2(t) = (1 - q_0) \cdot (-\alpha \ln(1 - qt)) + q_0,$$

математическое ожидание:

$$\mathsf{E}\xi = g_2'(1) = (1 - q_0) \cdot \frac{\alpha q}{1 - q},$$

дисперсия:

$$g_2''(t) = (1 - q_0) \cdot \frac{\alpha q^2}{(1 - qt)^2}$$

$$\mathsf{D}\xi = g_2''(1) + g_2'(1) - (g_2'(1))^2 = (1 - q_0) \cdot \frac{\alpha q^2}{(1 - q)^2} + (1 - q_0) \cdot$$

$$\cdot \frac{\alpha q}{1 - q} - (1 - q_0)^2 \cdot \frac{\alpha^2 q^2}{(1 - q)^2} = \alpha q (1 - q_0) \cdot \frac{1 - \alpha(1 - q_0)q}{(1 - q)^2},$$

рассеяние:

$$e\xi = \frac{\mathsf{D}\xi}{\mathsf{E}\xi} = \frac{1 - \alpha(1 - q_0)q}{1 - q}.$$

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

$$q = 1 - e^{q_0 - 1}.$$

Если мы посмотрим на данные в статье [4], 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 $q_0$ — в 0). Однако свойство смены знака логарифма рассеяния данного распределения будет использовано в дальнейшем при составлении сложных распределений на его основе.

## 1.3. Рассеяние сложных распределений

В статье Алексеевой [4] показано, что эмпирические данные должны иметь модель распределения, логарифм рассеяния которого имеет переменный знак, или, другими словами, рассеяние может быть больше и меньше 1. Однако многие дискретные распределения (биномиальное, пуассоновское, геометрическое) не обладают таким свойством. Было выдвинуто предположение, что логарифмическое распределение:

$$P(\xi = k) = \frac{\alpha q^k}{k}, \quad k = 1, 2, \ldots,$$

с параметром $q_0$, определяющим вероятность в 0, окажется подходящим, 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 $S_\tau$ — сумма случайного числа $\tau$ случайных величин $\xi_i$, одинаково распределённых и не зависимых между собой и от $\tau$. Тогда для её рассеяния справедлива формула:*

$$e(S_\tau) = \mathsf{E}\xi e(\tau) + e(\xi).$$

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

$$\mathsf{E}S_\tau = (h(g))'(1) = h'(g(1))g'(1) = \mathsf{E}\xi\mathsf{E}\tau,$$

и дисперсия:

$$\mathsf{D}S_\tau = (h(g))''(1) + (h(g))'(1) - \big((h(g))'(1)\big)^2 = (\mathsf{E}\xi)^2\,\mathsf{D}\tau + \mathsf{E}\tau\mathsf{D}\xi.$$

Поделив дисперсию на математическое ожидание как раз получим рассеяние:

$$e(S_\tau) = \frac{\mathsf{D}S_\tau}{\mathsf{E}S_\tau} = \frac{(\mathsf{E}\xi)^2\,\mathsf{D}\tau + \mathsf{E}\tau\mathsf{D}\xi}{\mathsf{E}\xi\mathsf{E}\tau} = \mathsf{E}\xi e(\tau) + e(\xi).$$

□

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 переменчивости знака логарифма рассеяния сложного распределения необходима переменчивость 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.

**Определение 3.** *Пусть $\xi_1, \xi_2, \ldots$ — независимые одинаково распределённые случайные величины, $\xi_i \sim Log(q), \forall i$, то есть удовлетворяют логарифмическому закону с параметром $q$, $\tau \sim Bin(n, p)$ — биномиальному закону с параметрами $n$ и $p$ и независима от $\xi_i, \forall i$.*

*Тогда случайная сумма $S_\tau = \xi_1 + \cdots + \xi_\tau$ будет случайной величиной, удовлетворяющей биномиально-логарифмическому распределению ($BinLog(n, p, q)$ или кратко «БЛР»).*

*Суперпозиция производящих функций биномиального и логарифмического распределения даст производящую функцию БЛР:*

$$h_1(t) = (-p\alpha \ln(1 - qt) + 1 - p)^n .$$

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

$$\mathsf{E}S_\tau = \mathsf{E}\xi_i \mathsf{E}\tau = \frac{np\alpha q}{1-q},$$

$$\mathsf{D}S_\tau = (\mathsf{E}\xi_i)^2 \mathsf{D}\tau + \mathsf{E}\tau \mathsf{D}\xi_i = np\alpha q \frac{1 - p\alpha q}{(1-q)^2},$$

$$e(S_\tau) = \mathsf{E}\xi_i e(\tau) + e(\xi_i) = \frac{1 - p\alpha q}{1-q}.$$

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

Для оценки параметров по методу максимального правдоподобия и проверки согласия с эмпирическими распределениями необходимо вычислить теоретические вероятности биномиально-логарифмического распределения.

### 2.1.1. Вероятности биномиально-логарифмического распределения

Как известно из свойств производящей функции [6]: если $h(t)$ — производящая функция дискретной случайной величины $\xi$, то её вероятности равны:

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

Тогда для производящей функции сложного распределения также справедливо:

$$P(S_\tau = k) = \frac{(h(g))^{(k)}(0)}{k!},$$

где $h(t)$ — производящая функция $\tau$, $g(t)$ — производящая функция $\xi$.

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:

$$g(0) = -\alpha \ln(1 - q \cdot 0) = 0,$$

то

$$h_1^{(k)}(0) = k! \cdot P(\tau = k), \quad \forall k = 0, 1, 2, \ldots$$

Тогда формула приобретает вид:

$$P(S_\tau = k) = \left(\frac{d^k}{dx^k}h(g)\right)(0) \cdot \frac{1}{k!} = \sum \frac{1}{m_1!m_2!\cdots m_k!}h^{(m_1+m_2+\cdots+m_k)}(0) \cdot \prod_{j=1}^{k}\left(\frac{g^{(j)}(0)}{j!}\right)^{m_j} =$$

$$= \sum \frac{1}{m_1! m_2! \cdots m_k!} (m_1 + m_2 + \cdots + m_k)! \cdot P(\tau = m_1 + m_2 + \cdots + m_k) \cdot \cdot \prod_{j=1}^{k} (P(\xi = j))^{m_j}.$$

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

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

Преобразуем, подставив вероятности соответствующих распределений:

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

Можно доказать общий вид для формулы вероятностей:

**Теорема 1.** *Вероятности биномиально-логарифмического распределения $S_\tau = \xi_1 + \cdots + \xi_\tau$, где $\xi_i$ распределены по логарифмическому закону с параметром $q$, $\tau$ — биномиальному с параметрами $n, p$, а*

$$h_1(t) = (-p\alpha \ln(1 - qt) + 1 - p)^n$$

*является производящей функцией, выражаются следующим образом:*

$$P(S_\tau = k) = \frac{1}{k!}(1-p)^{n-k} \cdot q^k \sum_{j=0}^{k} \frac{n!}{(n-j)!} s(k, j)(p\alpha)^j (1-p)^{k-j}, \quad k = 0, 1, \ldots \tag{2.1}$$

*где* $\alpha = \frac{-1}{\ln(1-q)}$, $s(k,j)$ *— число Стирлинга первого рода без знака [7], задающее число перестановок из* $k$ *элементов с* $j$ *циклами и имеющие следующую рекуррентную формулу:*

$$s(k+1,j) = s(k,j-1) + ks(k,j), 0 < j < k,$$

$$s(0,0) = 1, s(k,0) = 0 \text{ при } k > 0, s(0,j) = 0 \text{ при } j > 0.$$

*Доказательство.* По свойствам производящих функций вероятности выражаются через производные как:

$$P(S_\tau = k) = \frac{h_1^{(k)}(0)}{k!},$$

поэтому достаточно доказать формулу только для производной производящей функции.

Доказательство проведём методом математической индукции:

1. Для базы индукции подставим в производящую функцию 0:

   $$h_1(0) = (-p\alpha \ln(1 - q \cdot 0) + 1 - p)^n = (1-p)^n,$$

   что соответствует формуле (2.1) при $k = 0$.

2. Сделаем переход от $k$ к $k+1$.

   Для удобства обозначим

   $$m = -p\alpha \ln(1 - qt) + 1 - p$$

   и заметим, что

   $$m' = p\alpha \frac{q}{1-qt}, \quad m(0) = 1 - p.$$

   Уже доказано:

   $$h_1^{(k)}(t) = q^k \frac{m^{n-k}}{(1-qt)^k} \sum_{j=0}^{k} \frac{n!}{(n-j)!} s(k,j)(p\alpha)^j m^{k-j}$$

   (это выражение зависит от $t$, но достаточно в базе не подставлять $t = 0$, чтобы выражение было верным). Докажем, что

   $$h_1^{(k+1)}(t) = q^{k+1} \frac{m^{n-k-1}}{(1-qt)^{k+1}} \sum_{j=0}^{k+1} \frac{n!}{(n-j)!} s(k+1,j)(p\alpha)^j m^{k+1-j}.$$

Это следует из цепочки преобразований:

$$\left(q^k\frac{m^{n-k}}{(1-qt)^k}\sum_{j=0}^{k}\frac{n!}{(n-j)!}s(k,j)(p\alpha)^j m^{k-j}\right)' =$$
$$= q^k\frac{(n-k)m^{n-k-1}p\alpha\frac{q}{1-qt}(1-qt)^k + m^{n-k}kq(1-qt)^{k-1}}{(1-qt)^{2k}}\cdot$$
$$\cdot\sum_{j=0}^{k}\frac{n!}{(n-j)!}s(k,j)(p\alpha)^j m^{k-j}+$$
$$+q^k\frac{m^{n-k}}{(1-qt)^k}\sum_{j=0}^{k}\frac{n!}{(n-j)!}s(k,j)(p\alpha)^j(k-j)m^{k-j-1}p\alpha\frac{q}{1-qt} =$$
$$= q^{k+1}\frac{m^{n-k-1}}{(1-qt)^{k+1}}\left(((n-k)p\alpha+km)\sum_{j=0}^{k}\frac{n!}{(n-j)!}s(k,j)(p\alpha)^j m^{k-j}+\right.$$
$$\left.+\sum_{j=0}^{k}\frac{n!}{(n-j)!}s(k,j)(p\alpha)^{j+1}(k-j)m^{k-j}\right) =$$
$$= q^{k+1}\frac{m^{n-k-1}}{(1-qt)^{k+1}}\left(\sum_{j=0}^{k}\frac{n!}{(n-j)!}ks(k,j)(p\alpha)^j m^{k+1-j}+\right.$$
$$\left.+\sum_{j=0}^{k}\frac{n!}{(n-j)!}s(k,j)(p\alpha)^{j+1}(n-j)m^{k-j}\right) =$$
$$= q^{k+1}\frac{m^{n-k-1}}{(1-qt)^{k+1}}\left(\sum_{j=0}^{k}\frac{n!}{(n-j)!}ks(k,j)(p\alpha)^j m^{k+1-j}+\right.$$
$$\left.+\sum_{j=1}^{k+1}\frac{n!}{(n-j)!}s(k,j-1)(p\alpha)^j m^{k+1-j}\right) =$$
$$= q^{k+1}\frac{m^{n-k-1}}{(1-qt)^{k+1}}\left(ks(k,0)m^{k+1}+\right.$$
$$+\sum_{j=1}^{k}\frac{n!}{(n-j)!}(s(k,j-1)+ks(k,j))(p\alpha)^j m^{k+1-j}+$$
$$\left.+\frac{n!}{(n-k-1)!}s(k,k)(p\alpha)^{k+1}\right) =$$
$$= q^{k+1}\frac{m^{n-k-1}}{(1-qt)^{k+1}}\left(\sum_{j=1}^{k}\frac{n!}{(n-j)!}s(k+1,j)(p\alpha)^j m^{k+1-j}+\right.$$
$$\left.+\frac{n!}{(n-k-1)!}s(k+1,k+1)(p\alpha)^{k+1}\right) =$$
$$= q^{k+1}\frac{m^{n-k-1}}{(1-qt)^{k+1}}\sum_{j=0}^{k+1}\frac{n!}{(n-j)!}s(k+1,j)(p\alpha)^j m^{k+1-j}$$

Заметим, что $s(k,0) = 0, k > 0$ и $s(k,k) = 1\ \forall k \geqslant 0$, поэтому замена $s(k,0)$ на $s(k+1,0)$ (при $k = 0$ слагаемое обнуляет множитель $k$) и $s(k,k)$ на $s(k+1,k+1)$ оправдана.

Подставляем $t = 0$ и завершаем доказательство.

□

Числа Стирлинга первого рода 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 $s(4,0) = 0, s(4,1) = 6, s(4,2) = 11, s(4,3) = 6, s(4,4) = 1$.

Приведём небольшую часть таблицы чисел Стирлинга (таб. 2.1).

Таблица 2.1. Числа Стирлинга первого рода

| $k \setminus j$ | 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.

## 2.2. Логарифмически-биномиальное распределение

Мы рассматривали суперпозицию биномиального и логарифмического распределения, 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).

**Определение 4.** *Пусть $\xi_1, \xi_2, \ldots$ — независимые одинаково распределённые случайные величины, $\xi_i \sim Bin(n,p), \forall i$, то есть удовлетворяют биномиальному закону с параметрами $n$ и $p$, $\tau \sim Log(q)$ — логарифмическому закону с параметром $q$ и независима от $\xi_i, \forall i$.*

*Тогда случайная сумма $S_\tau = \xi_1 + \cdots + \xi_\tau$ будет случайной величиной, удовлетворяющей логарифмически-биномиальному распределению ($LogBin(q,n,p)$ или кратко «ЛБР»).*

*Суперпозиция производящих функций логарифмического и биномиального распределения даст производящую функцию ЛБР:*

$$h_2(t) = -\alpha \ln(1 - q(pt + 1 - p)^n).$$

Аналогично найдём для него основные характеристики:

$$\begin{aligned}
\mathsf{E}S_\tau &= \mathsf{E}\xi_i \mathsf{E}\tau = \frac{np\alpha q}{1-q}, \\
\mathsf{D}S_\tau &= (\mathsf{E}\xi_i)^2 \mathsf{D}\tau + \mathsf{E}\tau \mathsf{D}\xi_i = np\alpha q \frac{np + (1-p)(1-q) - np\alpha q}{(1-q)^2}, \\
e(S_\tau) &= \mathsf{E}\xi_i e(\tau) + e(\xi_i) = \frac{np + (1-p)(1-q) - np\alpha q}{1-q}.
\end{aligned}$$

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 логарифмически-биномиального распределения

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

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

Причём, если обозначить отдельно производящие функции для $\xi$ и $\tau$ как $h$ и $f$, то получим такие формулы для производных:

$$\begin{aligned}
h_2(t) =& f(h(t)) \\
h_2'(t) =& f'(h(t))h'(t) \\
h_2''(t) =& f''(h(t))(h'(t))^2 + f'(h(t))h''(t) \\
h_2'''(t) =& f'''(h(t))(h'(t))^3 + f''(h(t))2h'(t)h''(t) + \\
&+ f''(h(t))h'(t)h''(t) + f'(h(t))h'''(t) \\
h_2^{IV}(t) =& f^{IV}(h(t))(h'(t))^4 + 6f'''(h(t))(h'(t))^2h''(t) + 3f''(h(t))(h''(t))^2 + \\
&+ 4f''(h(t))h'(t)h'''(t) + f'(h(t))h^{IV}(t).
\end{aligned}$$

Тогда итоговые вероятности будут равны:

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

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

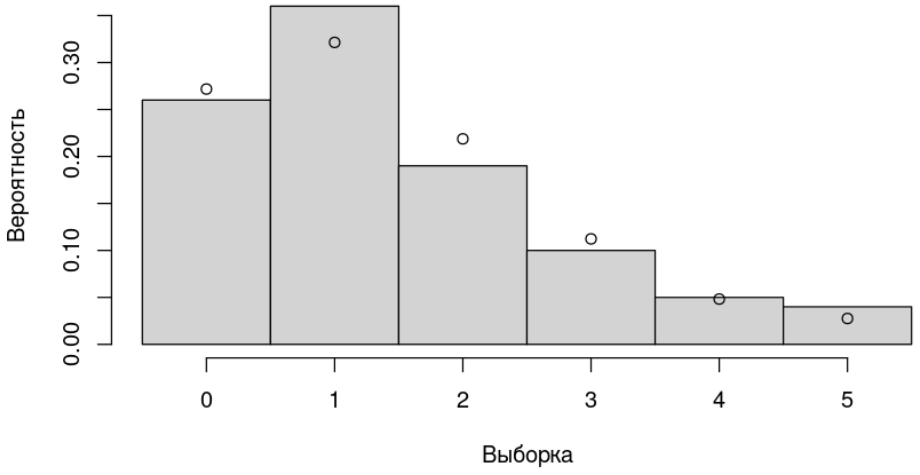


Рис. 2.1. Эмпирические частоты (столбики) и теоретические частоты (точки), вычисленные по методу максимального правдоподобия

## 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 больше 4 одинаковыми. Составим функцию правдоподобия:

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

Прологарифмируем:

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

Для оценки параметров методом правдоподобия необходимо искать минимум (максимум) функции, то есть применять математическую оптимизацию. С помощью метода моментов можно понизить размерность этой оценки.

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

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

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

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

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 $p$, $q$ и $n$ численными методами многомерной оптимизации (функция `optim` в R). Также для реальных оценок будем использовать численную оптимизацию по всем параметрам без метода моментов, чтобы увеличить точность.

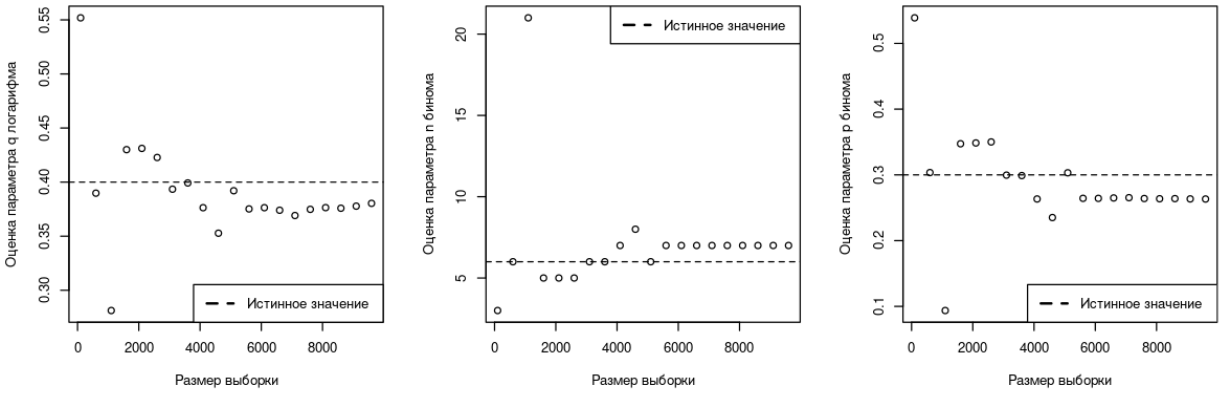


Рис. 2.2. Сходимость оценок параметров к истинному значению для биномиально-логарифмического распределения

Оценка параметра $n$ производилась методом целочисленного тернарного поиска (деление отрезка на три части и выбор из двух средних точек максимальной для сужения интервала на котором точно есть экстремум). Оценки максимального правдоподобия являются состоятельными, в чём можем убедиться моделированием (рис. 2.2 и 2.3),

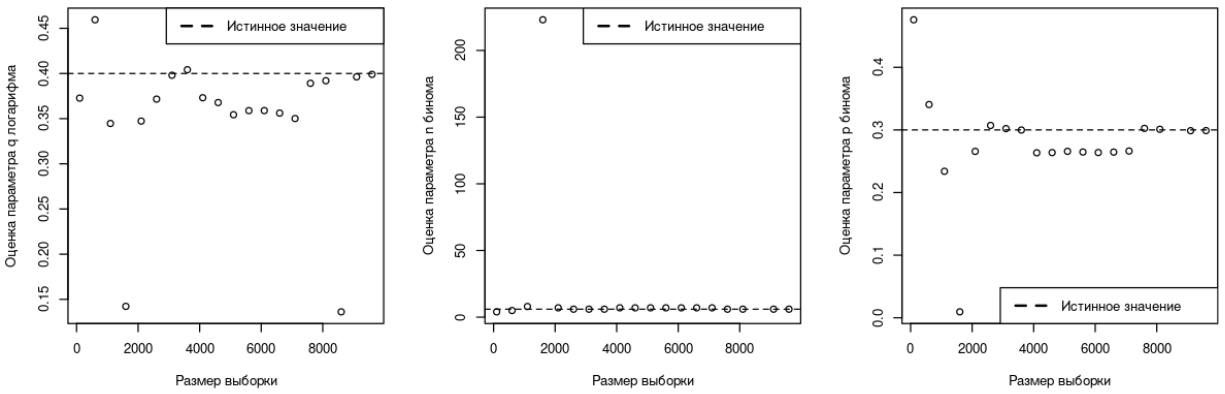


Рис. 2.3. Сходимость оценок параметров к истинному значению для логарифмически-биномиального распределения

хотя сходимость для данных распределений медленная из-за оценки трёх параметров одновременно, а для биномиально-логарифмического при небольшом размере выборки оценки получаются смещёнными.

Например, in vitro при 35 Гр получаем наилучшие оценки: $n = 866, q = 0.1799, p = 0.0015$. Значимость согласия по критерию $\chi^2$: $p - value = 0.26$. Результат представлен на рис. 2.1.

## 2.4. Применение биномиально-логарифмического и логарифмически-биномиального распределения в радиобиологии

При промежуточных результатах была выдвинута следующая интерпретация: $n$ — 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.

Однако мы применили логарифмически-биномиальное и биномиально-логарифмическое распределения к in vivo и in vitro данным одновременно. С помощью численной оптимизации трёхмерных функций правдоподобия для каждой выборки ядерных ано-

Оценки параметров и значимости критерия хи-квадрат биномиально-логарифмического распределения.

Таблица 2.2. In vivo

| Доза, Гр | n | q | p | **p-value** |
|---|---|---|---|---|
| 0 | 1 | 0.20 | 0.34 | **0.99** |
| 5 | 6 | 1.4e-6 | 0.11 | **0.63** |
| 10 | 2 | 0.21 | 0.37 | **0.62** |
| 15 | 2 | 0.24 | 0.50 | **0.15** |
| 20 | 82 | 0.24 | 0.02 | **0.03** |
| 25 | 4 | 0.21 | 0.24 | **0.60** |
| 30 | 997 | 0.16 | 1.6e-3 | **0.27** |
| 35 | 988 | 0.03 | 1.9e-3 | **0.03** |
| 40 | 983 | 0.01 | 2.2e-3 | **0.09** |
| 45 | 33 | 7.0e-6 | 0.07 | **0.02** |

Таблица 2.3. In vitro

| Доза, Гр | n | q | p | **p-value** |
|---|---|---|---|---|
| 0 | 10 | 3.9e-6 | 0.04 | **0.45** |
| 5 | 990 | 0.24 | 3.1e-3 | **0.84** |
| 10 | 960 | 0.08 | 5.7e-3 | **0.06** |
| 15 | 1 | 0.61 | 0.52 | **0.81** |
| 20 | 1 | 0.47 | 0.41 | **0.85** |
| 25 | 995 | 0.16 | 1.0e-3 | **0.28** |
| 30 | 2 | 0.47 | 0.40 | **0.77** |
| 35 | 866 | 0.18 | 1.5e-3 | **0.26** |
| 40 | 981 | 0.08 | 1.6e-3 | **0.24** |

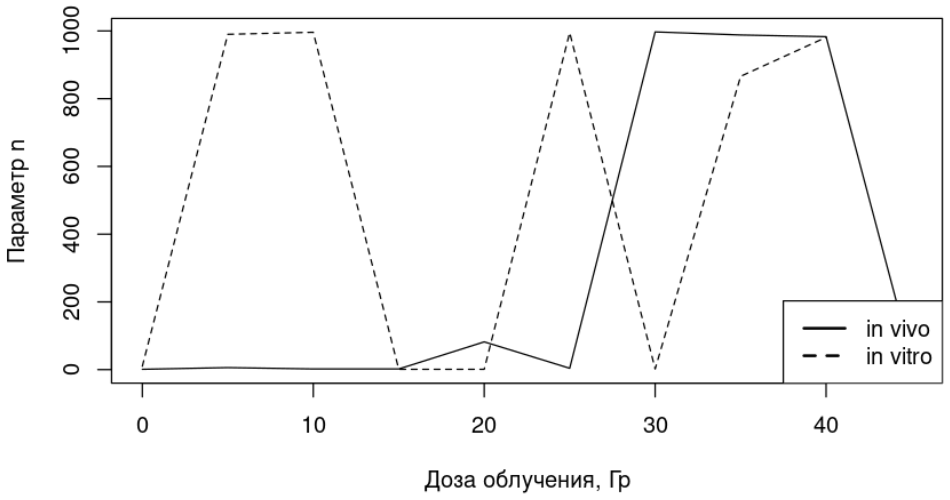


Рис. 2.4. Оценка параметра $n$ для биномиально-логарифмического распределения в зависимости от дозы облучения

малий у крыс получены результаты согласованности по критерию $\chi^2$, отражённые в таблицах 2.2, 2.3, 2.4, 2.5.

Оценки параметров и значимости критерия хи-квадрат логарифмически-биномиального распределения.

Таблица 2.4. In vivo

| Доза, Гр | n | q | p | **p-value** |
|---|---|---|---|---|
| 0 | 1 | 0.47 | 0.27 | **0.99** |
| 5 | 6 | 7.6e-7 | 0.11 | **0.63** |
| 10 | 20 | 4.3e-7 | 0.04 | **0.67** |
| 15 | 2 | 0.31 | 0.48 | **0.13** |
| 20 | 995 | 4.4e-6 | 1.7e-3 | **0.03** |
| 25 | 6 | 0.27 | 0.15 | **0.63** |
| 30 | 1000 | 0.32 | 1.5e-3 | **0.31** |
| 35 | 998 | 0.19 | 1.7e-3 | **0.11** |
| 40 | 1000 | 0.22 | 2.1e-3 | **0.16** |
| 45 | 998 | 0.18 | 2.3e-3 | **0.08** |

Таблица 2.5. In vitro

| Доза, Гр | n | q | p | **p-value** |
|---|---|---|---|---|
| 0 | 20 | 1.7e-6 | 0.02 | **0.52** |
| 5 | 994 | 0.80 | 1.4e-4 | **0.73** |
| 10 | 1000 | 1.1e-6 | 5.9e-4 | **0.03** |
| 15 | 2 | 0.80 | 0.18 | **0.72** |
| 20 | 981 | 0.54 | 3.9e-4 | **0.48** |
| 25 | 960 | 0.37 | 9.2e-4 | **0.22** |
| 30 | 2 | 0.69 | 0.31 | **0.61** |
| 35 | 8 | 0.52 | 0.12 | **0.52** |
| 40 | 8 | 0.43 | 0.16 | **0.73** |

Какой либо зависимости параметров от дозы облучения заметить не удаётся. Для примера покажем график для параметра $n$ в биномиально-логарифмическом распределении (рис. 2.4).

При этом для подавляющего большинства случаев биномиально-логарифмическое ($16/19 \cdot 100\% = 84\%$) и логарифмически-биномиальное распределения ($17/19 \cdot 100\% = 89\%$) дают согласованность на уровне $\alpha = 0.05$, что, с учётом небольшого числа случаев (всего 19) близко к теоретическим значениям.

Параметры $n$ и $p$ имеют большой разброс значений и часто $n$ устремляется в бесконечность (при оценке параметр ограничен 1000), а $p$ к нулю, что приводит нас к другому сложному распределению.

## 2.5. Логарифмически-пуассоновское распределение

Логарифмически-биномиальное распределение можно свести к распределению с меньшим количеством параметров. Для этого устремим параметр $n$ к бесконечности, а $p$ к нулю так, чтобы $n \cdot p = \lambda$. По теореме Пуассона, таким образом, из биномиального

распределения получается распределение Пуассона. Покажем это на примере производящей функции:

$$f(t) = (pt+1-p)^n = \left(1+\frac{1}{\frac{1-p}{pt}}\right)^n \cdot (1-p)^n = \left(1+\frac{1}{\frac{1-p}{pt}}\right)^{\frac{1-p}{pt} n \frac{pt}{1-p}} \cdot e^{n\ln(1-p)} \overset{(!)}{=}$$

$$\overset{(!)}{=} \left(\left(1+\frac{1}{\frac{1-p}{pt}}\right)^{\frac{1-p}{pt}}\right)^{n\frac{pt}{1-p}} \cdot e^{-np+no(p^2)} \xrightarrow[\substack{n\to\infty \\ n\cdot p=\lambda}]{p\to 0} e^{\lambda(t-1)},$$

где переход (!) сделан по разложению логарифма в районе нуля в ряд Тейлора, а при устремлении параметров использован замечательный предел.

Делаем аналогичный переход в производящей функции логарифмически-биномиального распределения и получаем новое распределение от 2-х параметров.

**Определение 5.** *Пусть $\xi_1, \xi_2, \ldots$ — независимые одинаково распределённые случайные величины, $\xi_i \sim Pois(\lambda), \forall i$, то есть удовлетворяют пуассоновскому закону с параметром $\lambda$, $\tau \sim Log(q)$ — логарифмическому закону с параметром $q$ и независима от $\xi_i, \forall i$.*

*Тогда случайная сумма $S_\tau = \xi_1 + \cdots + \xi_\tau$ будет случайной величиной, удовлетворяющей логарифмически-пуассоновскому распределению ($LogPois(q,\lambda)$ или кратко «ЛПР»).*

*Суперпозиция производящих функций логарифмического и пуассоновского распределения даст производящую функцию ЛПР:*

$$h_3(t) = -\alpha \ln(1 - qe^{\lambda(t-1)}).$$

Очевидным преимуществом данного распределения является вещественность всех его параметров, а также, в принципе, их количество: два против трёх. Однако было утеряно свойство переменчивости знака логарифма рассеяния, что следует, в том числе, из леммы 1

Обозначим простейшие характеристики данного распределения:

$$\mathsf{E}S_\tau = \mathsf{E}\xi_i \mathsf{E}\tau = \frac{\lambda\alpha q}{1-q},$$

$$\mathsf{D}S_\tau = (\mathsf{E}\xi_i)^2 \mathsf{D}\tau + \mathsf{E}\tau \mathsf{D}\xi_i = \lambda\alpha q \frac{\lambda + (1-q) - \lambda\alpha q}{(1-q)^2},$$

$$e(S_\tau) = \mathsf{E}\xi_i e(\tau) + e(\xi_i) = \frac{\lambda + (1-q) - \lambda\alpha q}{1-q}.$$

Ещё раз убеждаемся в том, что рассеяние больше 1:

$$\frac{\lambda q + \lambda \ln(1-q) + (1-q)\ln(1-q)}{\ln(1-q)(1-q)} = 1$$

$$\lambda(q + \ln(1-q)) = 0,$$

А это справедливо только при $\lambda = 0$ или $q = 0$.

### 2.5.1. Вероятности логарифмически-пуассоновского распределения

Вероятности находим аналогично биномиально-логарифмическому и логарифмически-биномиальному распределениям через формулу для производящих функций.

**Теорема 2.** *Вероятности логарифмически-пуассоновского распределения с параметрами $\lambda > 0$ и $q \in (0,1)$ и производящей функцией:*

$$h_3(t) = -\alpha \ln\left(1 - qe^{\lambda(t-1)}\right),$$

*равны при $k = 0$:*

$$P(S_\tau = 0) = -\alpha \ln\left(1 - qe^{-\lambda}\right),$$

*а при $k = 1, 2 \ldots$:*

$$P(S_\tau = k) = \frac{\alpha}{k!}\lambda^k \sum_{j=1}^{k} (j-1)! S(k,j) \left(\frac{qe^{-\lambda}}{1 - qe^{-\lambda}}\right)^j, \tag{2.2}$$

*где $S(k,j)$ — числа Стирлинга второго рода [7], имеющие следующую рекуррентную формулу:*

$$S(k,j) = S(k-1, j-1) + j \cdot S(k-1, j), \quad 0 < j \leqslant k,$$

$$S(0,0) = 1, S(k,0) = 0 \text{ при } k > 0, S(k,j) = 0 \text{ при } j > k.$$

*Доказательство.* Вероятность для $k = 0$ получается простой подстановкой нуля в производящую функцию.

Для доказательство остальных значений $k$ воспользуемся методом математической индукции:

1. Докажем для $k = 1$:

$$P(S_\tau = 1) = \frac{1}{1!}\left(h_3(t)\right)'\Big|_{t=0} = \left(-\alpha \ln\left(1 - qe^{\lambda(t-1)}\right)\right)'\Big|_{t=0} = -\alpha \frac{-q\lambda e^{\lambda(t-1)}}{1 - qe^{\lambda(t-1)}}\Big|_{t=0} =$$

$$= \alpha\lambda \frac{qe^{-\lambda}}{1 - qe^{-\lambda}},$$

что соответствует формуле (2.2) при подстановке в неё $k = 1$.

2. Докажем переход от $k$ к $k+1$.

Для этого обозначим $qe^{\lambda(t-1)} = m$ и заметим, что $m' = \left(qe^{\lambda(t-1)}\right)' = \lambda qe^{\lambda(t-1)} = \lambda m$.

Уже доказано, что

$$h_3^{(k)}(t) = \alpha\lambda^k \sum_{j=1}^{k}(j-1)!S(k,j)\left(\frac{m}{1-m}\right)^j$$

(это выражение зависит от $t$, но достаточно в первом пункте при $k=1$ не подставлять $t=0$, чтобы база математической индукции была верна). Докажем, что

$$h_3^{(k+1)}(t) = \alpha\lambda^{k+1} \sum_{j=1}^{k+1}(j-1)!S(k+1,j)\left(\frac{m}{1-m}\right)^j.$$

Это следует из цепочки равенств:

$$\left(\lambda^k \sum_{j=1}^{k}(j-1)!S(k,j)\left(\frac{m}{1-m}\right)^j\right)' =$$

$$= \lambda^k \sum_{j=1}^{k}(j-1)!S(k,j)j\left(\frac{m}{1-m}\right)^{j-1} \cdot \left(\frac{\lambda m}{1-m} + \frac{\lambda m^2}{(1-m)^2}\right) =$$

$$= \lambda^{k+1}\left(\sum_{j=1}^{k} j!S(k,j)\left(\frac{m}{1-m}\right)^j + \sum_{j=1}^{k} j!S(k,j)\left(\frac{m}{1-m}\right)^{j+1}\right) =$$

$$= \lambda^{k+1}\left(\sum_{j=1}^{k} j!S(k,j)\left(\frac{m}{1-m}\right)^j + \sum_{j=2}^{k+1}(j-1)!S(k,j-1)\left(\frac{m}{1-m}\right)^j\right) =$$

$$= \lambda^{k+1}\left(S(k,1)\frac{m}{1-m} + \right.$$

$$\left. + \sum_{j=2}^{k}\left(j!S(k,j)\left(\frac{m}{1-m}\right)^j + (j-1)!S(k,j-1)\left(\frac{m}{1-m}\right)^j\right) + \right.$$

$$\left. + k!S(k,k)\left(\frac{m}{1-m}\right)^{k+1}\right) =$$

$$= \lambda^{k+1}\left(S(k+1,1)\frac{m}{1-m} + \sum_{j=2}^{k}(j-1)!\left(jS(k,j) + S(k,j-1)\right)\left(\frac{m}{1-m}\right)^j + \right.$$

$$\left. + k!s(k+1,k+1)\left(\frac{m}{1-m}\right)^{k+1}\right) =$$

$$= \lambda^{k+1}\left(S(k+1,1)\frac{m}{1-m} + \sum_{j=2}^{k}(j-1)!S(k+1,j)\left(\frac{m}{1-m}\right)^j + \right.$$

$$\left. + k!s(k+1,k+1)\left(\frac{m}{1-m}\right)^{k+1}\right) =$$

$$= \lambda^{k+1} \sum_{j=1}^{k+1} (j-1)! S(k+1, j) \left( \frac{m}{1-m} \right)^j.$$

Заметим, что $S(l, 1) = 1 \ \forall l > 0$ и $S(m, m) = 1 \ \forall m \geqslant 0$, поэтому замены $S(k, 1)$ на $S(k+1, 1)$ и $S(k, k)$ на $S(k+1, k+1)$ легитимные.

Коэффициент $\alpha$ не влияет на дифференцирование, а также $qe^{-\lambda} = m|_{t=0}$, поэтому переход доказан.

□

Полученная формула имеет связь с числами Стирлинга второго рода, которые обозначают количество неупорядоченных разбиений $k$-элементного множества на $j$ непустых подмножеств. В разделе 2.1.1 была показана формула для вероятности через числа Стирлинга первого рода, что говорит о том, что данные числа могут иметь большую важность для различных классов сложных распределений.

Явная формула для чисел Стирлинга:

$$S(k, j) = \frac{1}{j!} \sum_{i=0}^{j} (-1)^{j+i} C_j^i i^k.$$

Приведём небольшую часть таблицы чисел Стирлинга (таб. 2.6).

Таблица 2.6. Числа Стирлинга второго рода

| $k \setminus j$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 1 | | | | | | |
| 1 | 0 | 1 | | | | | |
| 2 | 0 | 1 | 1 | | | | |
| 3 | 0 | 1 | 3 | 1 | | | |
| 4 | 0 | 1 | 7 | 6 | 1 | | |
| 5 | 0 | 1 | 15 | 25 | 10 | 1 | |
| 6 | 0 | 1 | 31 | 90 | 65 | 15 | 1 |

Для логарифмически-пуассоновского распределения можно получить формулу другого вида с другим классом чисел.

**Теорема 3.** *В условиях теоремы 2 вероятности логарифмически-пуассоновского распределения могут быть выражены следующим образом:*

1. *Для $k = 0, 1$ формула аналогична теореме 2;*

2. *Для $k = 2, ...$:*

$$P(S_\tau = k) = \frac{\alpha}{k!} \frac{\lambda^k q e^{-\lambda}}{(1 - qe^{-\lambda})^k} \sum_{j=0}^{k-2} E(k-1, j) \left(qe^{-\lambda}\right)^j,$$

*где $E(k, j)$ — числа Эйлера первого рода [7]. Они имеют следующую рекуррентную формулу:*

$$E(k, j) = (j+1) \cdot E(k-1, j) + (k-j) \cdot E(k-1, j-1), 0 < j < k-1,$$

$$E(k, 0) = 1 \text{ при } k \geqslant 0, E(k, j) = 0 \text{ при } j \geqslant k > 0.$$

*Доказательство.* Формулы для $k = 0, 1$ получены аналогично теореме 2.

Проведём математическую индукцию для остальных значений $k$. Сразу применим замену $m = qe^{\lambda(t-1)}$.

1. База индукции:

$$P(S_\tau = 2) = \frac{1}{2!} \left(h_3(t)\right)'' \Big|_{t=0} = \left(-\alpha \ln(1-m)\right)'' \big|_{t=0} = \frac{1}{2} \left(-\alpha \frac{-\lambda m}{1-m}\right)' \Big|_{t=0} =$$
$$= \frac{\alpha}{2} \lambda \frac{\lambda m (1-m) - m(-\lambda m)}{(1-m)^2} \Big|_{t=0} = \frac{\alpha}{2} \frac{\lambda^2 q e^{-\lambda}}{(1 - qe^{-\lambda})^2},$$

2. Докажем переход от $k$ к $k+1$.

   Уже доказано, что

$$h_3^{(k)}(t) = \alpha \frac{\lambda^k q e^{-\lambda}}{(1-m)^k} \sum_{j=0}^{k-2} E(k-1, j) m^j$$

   (это выражение зависит от $t$, но достаточно в первом пункте при $k = 1$ не подставлять $t = 0$, чтобы база математической индукции была верна). Докажем, что

$$h_3^{(k+1)}(t) = \alpha \frac{\lambda^{k+1} m}{(1-m)^{k+1}} \sum_{j=0}^{k-1} E(k, j) m^j.$$

Это следует из следующей цепочки:

$$\left(\frac{\lambda^k}{(1-m)^k}\sum_{j=0}^{k-2}E(k-1,j)m^{j+1}\right)' =$$
$$= \frac{km\lambda^{k+1}}{(1-m)^{k+1}}\sum_{j=0}^{k-2}E(k-1,j)m^{j+1} + \frac{(1-m)\lambda^k}{(1-m)^{k+1}}\sum_{j=0}^{k-2}E(k-1,j)\lambda(j+1)m^{j+1} =$$
$$= \frac{\lambda^{k+1}}{(1-m)^{k+1}}\cdot\left(\sum_{j=0}^{k-2}E(k-1,j)km^{j+2} + \sum_{j=0}^{k-2}E(k-1,j)(j+1)m^{j+1} - \right.$$
$$\left. - \sum_{j=0}^{k-2}E(k-1,j)(j+1)m^{j+2}\right) =$$
$$= \frac{\lambda^{k+1}}{(1-m)^{k+1}}\cdot\left(E(k-1,0)m + \right.$$
$$+ \sum_{j=1}^{k-2}\left(E(k-1,j)(j+1) + E(k-1,j-1)(k-j)\right)m^{j+1} +$$
$$\left. + E(k-1,k-2)m^k\right) =$$
$$= \frac{\lambda^{k+1}}{(1-m)^{k+1}}\left(E(k,0)m + \sum_{j=1}^{k-2}E(k,j)m^j + E(k,k-1)m^k\right) =$$
$$= \frac{\lambda^{k+1}}{(1-m)^{k+1}}\sum_{j=0}^{k-1}E(k,j)m^{j+1} = \frac{\lambda^{k+1}m}{(1-m)^{k+1}}\sum_{j=0}^{k-1}E(k,j)m^j.$$

Заметим, что $E(l,0) = 1\ \forall l \geqslant 0$ и $E(m,m-1) = 1\ \forall m \geqslant 1$, поэтому замены $E(k-1,0)$ на $E(k,0)$ и $E(k-1,k-2)$ на $E(k,k-1)$ легитимные.

Подставляем $t = 0$, умножаем на $\alpha/(k+1)!$ и получаем нужную формулу.

□

Числа Эйлера представляют собой количество перестановок порядка $k$ с $j$ подъёмами, то есть в перестановке $\pi = (\pi_1, \ldots, \pi_k)$ всего $j$ индексов $i$ таких, что $\pi_i < \pi_{j+1}$. Приведём небольшую часть таблицы чисел Эйлера (таб. 2.7).

Для них можно привести явную формулу:

$$E(k,j) = \sum_{i=0}^{j}C_{k+1}^i(-1)^i(j+1-i)^k.$$

Таблица 2.7. Числа Эйлера первого рода

| $k \setminus j$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 1 | | | | | |
| 1 | 1 | | | | | |
| 2 | 1 | 1 | | | | |
| 3 | 1 | 4 | 1 | | | |
| 4 | 1 | 11 | 11 | 1 | | |
| 5 | 1 | 26 | 66 | 26 | 1 | |
| 6 | 1 | 57 | 302 | 302 | 57 | 1 |

Эти числа также возникают в следующих бесконечных функциональных рядах:

$$\sum_{i=1}^{\infty} i^k x^i = \frac{x}{(1-x)^{k+1}} \sum_{j=0}^{k-1} E(k,j) x^j.$$

Поэтому можно убедиться в справедливости формул вероятностей из теоремы 3. Используем формулу полной вероятности, чтобы найти вероятность случайной суммы случайных величин:

$$\begin{aligned} P(S_\tau = k) =& \sum_{j=1}^{\infty} P(S_\tau = k | \tau = j) \cdot P(\tau = j) = \sum_{j=1}^{\infty} -\frac{1}{\ln(1-q)} \frac{q^j}{j} \cdot P(\xi_1 + \cdots + \xi_j = k) = \\ =& \sum_{j=1}^{\infty} -\frac{1}{\ln(1-q)} \frac{q^j}{j} \frac{(j\lambda)^k}{k!} e^{-j\lambda} = \frac{\alpha}{k!} \lambda^k \sum_{j=1}^{\infty} j^{k-1} \left(qe^{-\lambda}\right)^j = \\ =& \frac{\alpha}{k!} \frac{\lambda^k q e^{-\lambda}}{(1-qe^{-\lambda})^k} \sum_{j=0}^{k-2} E(k-1,j) \left(qe^{-\lambda}\right)^j. \end{aligned}$$

Компьютерное вычисление вероятностей показало эквивалентность этих трёх формул с погрешностью на уровне машинного нуля.

### 2.5.2. Оценка параметров логарифмически-пуассоновского распределения

Оценка параметров $\lambda$ и $q$ проводилась методом максимального правдоподобия численно на компьютере с использованием функции `optim` на языке `R`.

Моделируя выборку из $n$ индивидов для известных параметров можно проверить состоятельность такой оценки.

По рис. 2.5 и 2.6 видим, что состоятельность имеет место для данных оценок.

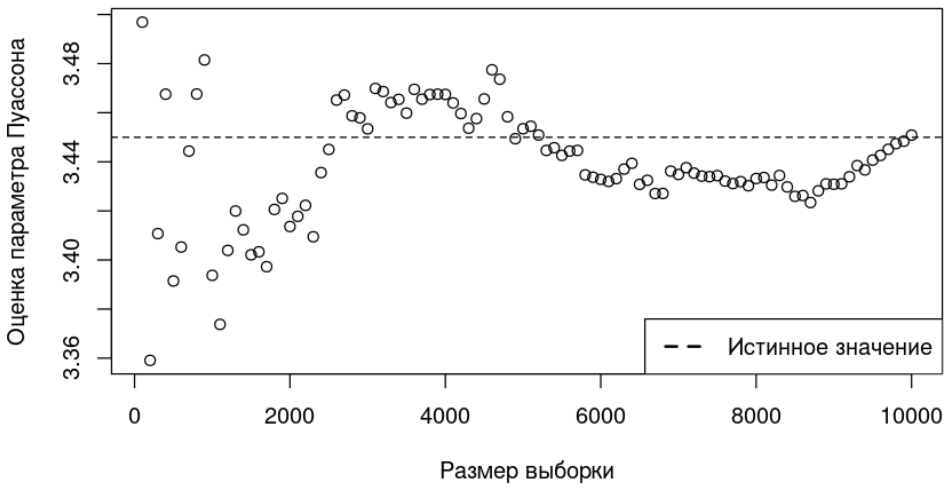


Рис. 2.5. Сходимость оценки параметра $\lambda$ к истинному значению для логарифмически-пуассоновского распределения

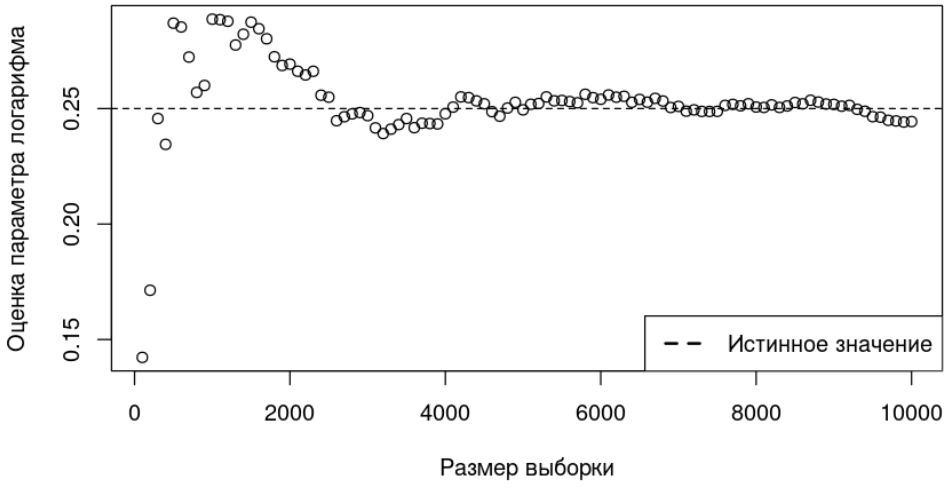


Рис. 2.6. Сходимость оценки параметра $q$ к истинному значению для логарифмически-пуассоновского распределения

### 2.5.3. Логарифмически-пуассоновское распределение в радиобиологии

Как уже было сказано выше, логарифмически-биномиальное распределение давало хорошую согласованность для радиобиологических данных в случае in vivo и in vitro, од-

Оценки параметров и значимости критерия хи-квадрат по данным in vivo и in vitro для логарифмически-пуассоновского распределения.

Таблица 2.8. In vivo

| Гр | $\lambda$ | $q$ | **p-v** |
|---|---|---|---|
| 0 | 0.38 | 5.9e-7 | **0.13** |
| 5 | 0.67 | 3.3e-7 | **0.81** |
| 10 | 0.83 | 6.7e-7 | **0.21** |
| 15 | 1.15 | 5.3e-7 | **0.01** |
| 20 | 1.71 | 1e-5 | **0.03** |
| 25 | 1.04 | 0.07 | **0.55** |
| 30 | 1.48 | 0.33 | **0.33** |
| 35 | 1.75 | 0.19 | **0.11** |
| 40 | 2.05 | 0.22 | **0.16** |
| 45 | 2.36 | 0.18 | **0.08** |

Таблица 2.9. In vitro

| Гр | $\lambda$ | $q$ | **p-v** |
|---|---|---|---|
| 0 | 0.39 | 3.7e-6 | **0.59** |
| 5 | 0.14 | 0.80 | **0.73** |
| 10 | 0.59 | 1.1e-7 | **0.03** |
| 15 | 0.41 | 0.76 | **0.36** |
| 20 | 0.37 | 0.56 | **0.45** |
| 25 | 0.88 | 0.37 | **0.22** |
| 30 | 0.78 | 0.53 | **0.26** |
| 35 | 1.10 | 0.43 | **0.43** |
| 40 | 1.46 | 0.29 | **0.38** |

нако его функция максимального правдоподобия имела неоднозначный минимум, представленный «оврагом», на котором параметры $n$ и $p$ уходили в бесконечность и к нулю, соответственно. Поэтому логично, что согласованность логарифмически-пуассоновского распределения будет схожа (таблицы 2.8 и 2.9), однако из-за потери переменчивости знака логарифма рассеяния появился один дополнительный случай (15 Гр. in vivo), где p-value меньше уровня значимости.

### 2.5.4. Интерпретация компонент логарифмически-пуассоновского распределения

Наблюдаемое количество аномалий определяется в целом двумя факторами: их исходной распространенностью и интенсивностью их образования в процессе митоза. За увеличение исходной распространенности отвечает параметр $q$ логарифмического распределения, а за интенсивность образования аномалий в процессе митоза параметр $\lambda$ пуассоновского распределения. Поскольку распределения суммы и самих случайных величин однопараметрические, при интерпретации параметров можно опираться на их

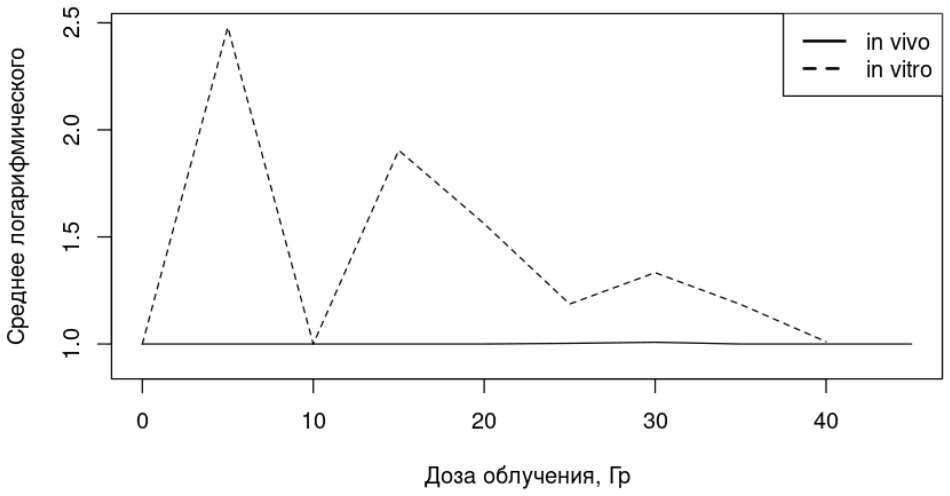


Рис. 2.7. Среднее логарифма в логарифмически-пуассоновском распределении в зависимости от дозы облучения

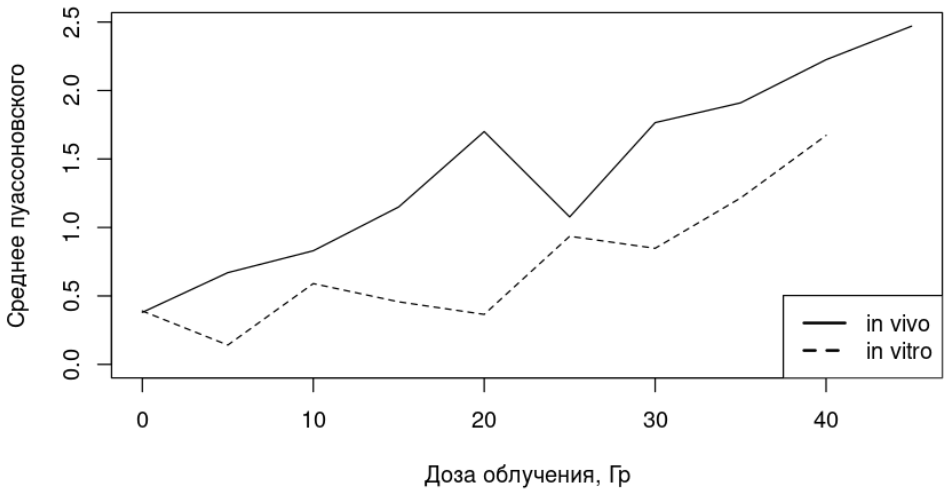


Рис. 2.8. Среднее Пуассона в логарифмически-пуассоновском распределении в зависимости от дозы облучения

средние значения.

Динамика оценок параметра $\lambda$, соответственно средних значений, свидетельствует о положительной линейной зависимости их от дозы облучения, что очевидно, и о значимо меньших значениях в эксперименте in vitro (рис. 2.8), так как выжившие клетки

очевидно обладают большим иммунитетом.

Что касается исходной распространенности аномалий, то, согласно динамике оценок параметра $q$ в зависимости от дозы облучения (рис. 2.7), зависимости от дозы облучения нет, а в эксперименте in vivo базовая распространенность практически не выражена и существенно меньше, чем в эксперименте in vitro. Это объясняется тем, что от дозы облучения зависит только количество выживших клеток, а не распространенность их аномалий, которая, судя по графику, очень вариабельна, но существенно выше базовой распространенности до начала эксперимента.

# Глава 3

# Применение к анализу встречаемости слов

## 3.1. 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 $\omega_j$ встречается в $i$-ой главе $x_i^{(j)}$ раз. Вопрос: какому распределению удовлетворяет выборка $(x_1^{(j)}, \ldots, x_n^{(j)})$?

В работе [5] 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.

## 3.2. Применение различных моделей 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.

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.

### 3.2.1. Нормированные числа Стирлинга первого рода

Для борьбы с вычислительными проблемами введём понятие нормированных чисел Стирлинга первого рода.

**Определение 6.** *Нормированными числами Стирлинга первого рода будем называть числа, получаемые следующим образом:*

$$l(k,j) = \frac{s(k,j)}{k!}.$$

То есть они являются числами Стирлинга первого рода, делёнными на факториал числа элементов перестановок (то есть $k$).

Выведем рекуррентную формулу для полученных чисел:

$$\begin{aligned} s(k,j) =& s(k-1,j-1) + (k-1)\cdot s(k-1,j) \quad | : (k!) \\ \frac{s(k,j)}{k!} =& \frac{s(k-1,j-1)}{k!} + \frac{(k-1)\cdot s(k-1,j)}{k!} \\ l(k,j) =& \frac{1}{k} l(k-1,j-1) + \frac{(k-1)}{k} l(k-1,j). \end{aligned}$$

Так как числа Стирлинга первого рода обозначают количество перестановок длины $k$ с $j$ циклами, то сумма по строчкам должна быть равна $k!$, а значит, сумма по строчкам у нормированных чисел Стирлинга равна 1. Приведём часть таблицы (таб. 3.1).

### 3.2.2. Оценки и согласованность

Изначально в качестве моделей встречаемости слов были взяты отрицательный бином и биномиально-логарифмическое распределение, однако основная часть слов, для которой не было достигнуто согласование — это слова, распределение которых имеет «тяжёлый» правый хвост, выражающийся в элементах выборки, которые имеют значение встречаемости сильно больше остальных. Пример такого слова представлен на

Таблица 3.1. Нормированные числа Стирлинга первого рода

| $k \setminus j$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 1 | | | | | | |
| 1 | 0 | 1 | | | | | |
| 2 | 0 | 1/2 | 1/2 | | | | |
| 3 | 0 | 1/3 | 1/2 | 1/6 | | | |
| 4 | 0 | 1/4 | 11/24 | 1/4 | 1/24 | | |
| 5 | 0 | 1/5 | 5/12 | 7/24 | 1/12 | 1/120 | |
| 6 | 0 | 1/6 | 137/360 | 5/16 | 17/144 | 1/48 | 1/720 |

Таблица 3.2. Разбиение слов по классам согласованности

| Распределение | Согласованные слова |
|---|---|
| ОБР | 1 |
| БЛР | 7 |
| Сумма | 6 |
| ОБР и БЛР | 25 |
| ОБР и сумма | 4 |
| БЛР и сумма | 18 |
| Все | 884 |
| Итого | 945 |

рис. 3.1. В предположении такое может произойти, если слово употребляется в нескольких значениях. Поэтому есть желание проверить сумму случайных величин, распределённых по ОБР. Согласованность проверялась по критерию $\chi^2$ с оцениванием параметров по методу максимального правдоподобия.

На уровне значимости $\alpha = 0.05$ было посчитано p-value для каждого распределения и все слова были разбиты на классы по согласованности с конкретными распределениями (таблица 3.2).

Заметим, что отрицательно-биномиальное распределение даёт согласование для $1 + 25 + 4 + 884 = 914$ слов, а биномиально-логарифмическое и сумма двух отрицательно-биномиальных добавляет ещё 31 слово, то есть несильно расширяют класс со-

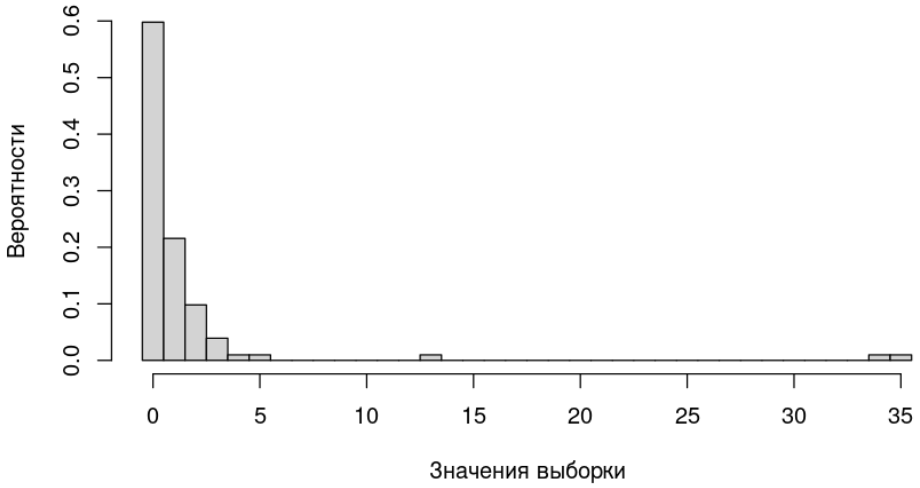


Рис. 3.1. Гистограмма распределения слова "coat"

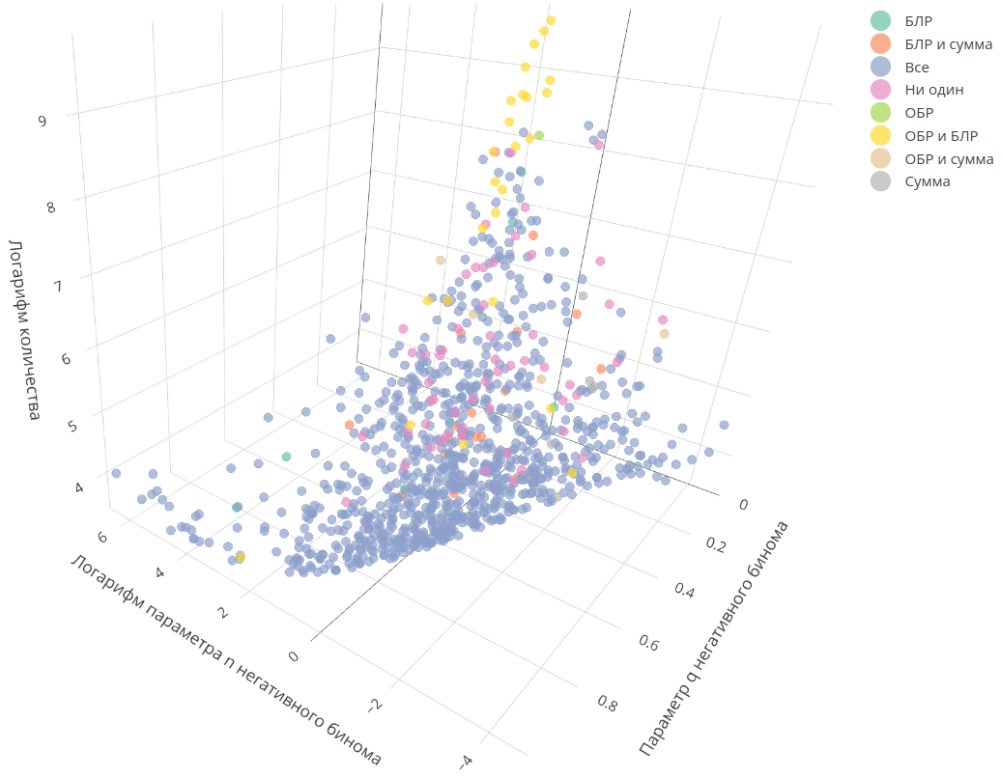


Рис. 3.2. Точечный график слов по параметрам ОБР и встречаемости, разбитые по классам.

гласованных слов, однако вместе с ними получается всего 945 согласованных слов, что почти равняется $\gamma \cdot 100\% = (1 - \alpha) \cdot 100\% = 95\%$ слов, то есть можно предположить, что набор данных моделей покрывает всевозможные вариации встречаемости слов в

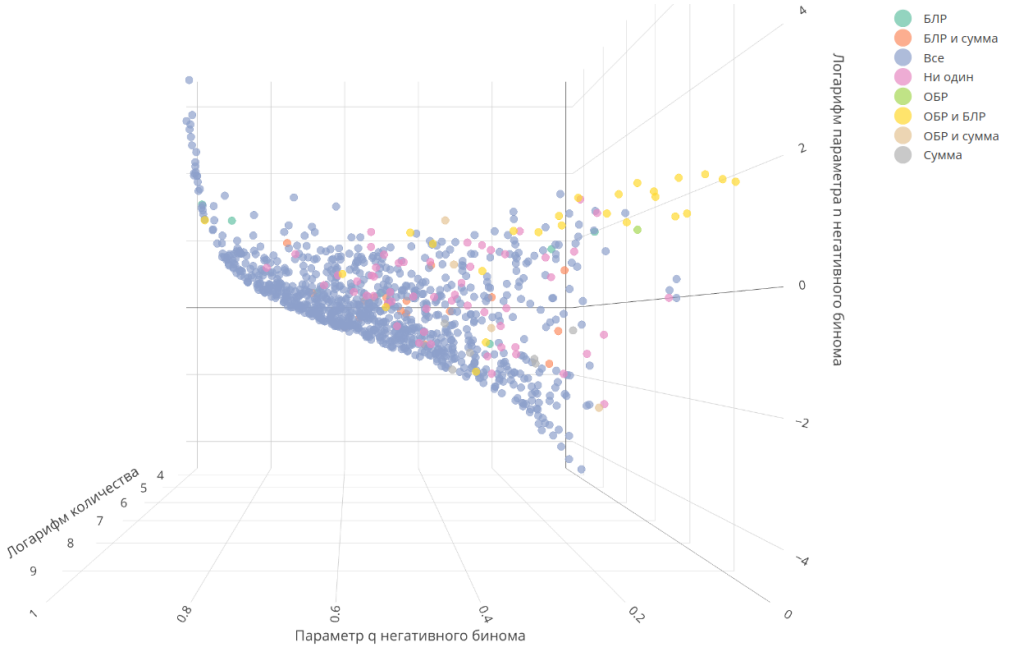


Рис. 3.3. Точечный график слов по параметрам ОБР и встречаемости, разбитые по классам.

текстах.

На рис. 3.2 представлен график оценок параметров и общей встречаемости слов. По оси $x$ откладывается значение оценки параметра $q$ в отрицательно-биномиальном распределении, по $y$ — логарифм параметра $n$, а по $z$ — общая встречаемость слова во всём тексте. Видно, что почти все слова располагаются на некоторой двухмерной поверхности, причём если мы посмотрим сверху на этот график, то увидим для каждого среза по оси $z$ некоторую кривую, ограничивающую все точки. К тому же, сдвиг этой кривой в зависимости от встречаемости обусловлен средним значением встречаемости.

# 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 [4]. 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.

Большой разброс значений оценок параметров показал возможную избыточность модели, поэтому было рассмотрено логарифмически-пуассоновское распределение, как сужение логарифмически-биномиального при стремлении параметра $n$ к бесконечности, а параметра $p$ к нулю. Для него были также получены простые характеристики, однако была потеряна переменчивость знака логарифма рассеяния, что несильно сказалось на применимости к радиобиологическим данным. Были найдены три вариации формул для вероятностей этого распределения, одна из которых связана с числами Стирлинга второго рода, а другая — с числами Эйлера первого рода. Применение к радиобиологическим данным дало почти ту же согласованность, что и логарифмически-биномиальное, причём удалось достигнуть более явной интерпретации компонент.

К встречаемости слов в тексте были применены обратно-биномиальное распределе-

ние и его обобщение биномиально-логарифмическое. В предположении многозначности некоторых слов, которая выражается в « тяжёлых»  хвостах добавлена модель суммы двух отрицательных биномов. В совокупности трёх моделей получено согласование при $\alpha = 0.05$ для 945 слов из 1000, что близко к теоретическому значению $\gamma = 0.95$. При этом за счёт биномиально-логарифмического распределения и суммы отрицательных биномов класс согласованных слов расширен лишь на 3.1%, что говорит о специфичности применимости данных моделей.

# Список литературы


1. Ватутин В. А. Лекционные курсы НОЦ. — Вып. 8: Ветвящиеся процессы и их применения изд. — Москва : МИАН, 2008. — ISBN: 5-98419-024-9.
2. Алексеева Н. П. Анализ медико-биологических систем. Реципрокность, эргодичность, синонимия. — Санкт-Петербург : Изд-во СПбГУ, 2013. — ISBN: 978-5-288-05347-4.
3. Бойков А. В. Модель Крамера–Лундберга со стохастическими премиями // Теория вероятн. и ее примен. — 2002. — Т. 47, № 3. — С. 549–553.
4. Динамика роста числа ядерных аномалий рабдомиосаркомы RA-23 при увеличении дозы острого редкоионизирующего облучения. Исследование на основе модели реинтрантно-биномиального распределения / Алексеева Н. П., Алексеев А. О., Вахтин Ю. Б., Кравцов В. Ю., Кузоватов С. Н. и Скорикова Т. И. // Цитология. — 2008. — № 6. — С. 528–534.
5. Alexeeva N. Sotov A. The Negative Binomial Model of Word Usage // Electronic Journal of Applied Statistical Analysis. — 2013. — Vol. 6, no. 1.
6. Феллер В. Введение в теорию вероятностей и её приложения. В 2 т. — Москва : Мир, 1952. — Т. 1.
7. Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики. — Москва : Мир, 1998. — ISBN: 5-03-001793-3.

# Приложение А

# Код

## А.1. Моделирование в R

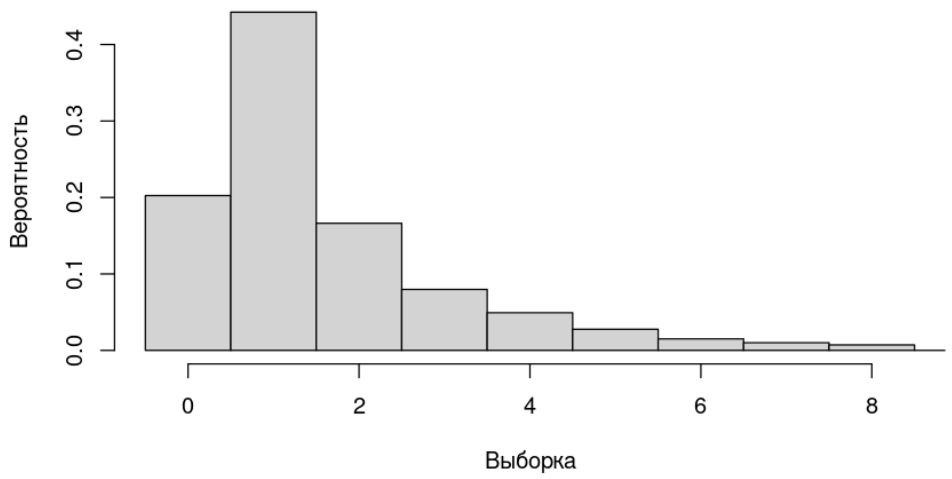


Рис. А.1. Промоделированное логарифмическое распределение с параметром. $p_0 = 0.2$ и $p = 0.75$

Анализ эмпирических данных проводился с использованием интерпретируемого языка программирования `R`, который изначально создан для статистического анализа данных. Поэтому его базовый функционал подразумевает встроенные функции моделирования случайных величин таких, как нормальное, пуассоновское, биномиальное, равномерное и пр. Однако, функция моделирования логарифмического распределения не реализована (справедливости ради, существуют пакеты с реализацией данного распределения), поэтому реализуем её сами.

Можно использовать несколько подходов в данном вопросе, однако, учитывая простой вид вероятностей, а также дискретность случайной величины, применим метод обратных функций.

Для этого сначала напишем функцию, которая по параметру будет вычислять таблицу функции распределения, накапливая вероятности:

```
log_prepering <- function(q = 0.5){
  tdistr <- (-q) / log(1 - q)
  sum_distr <- tdistr
  sum <- c(sum_distr)
  k <- 2

  while(tdistr > 1e-10){
    tdistr <- tdistr * q / k * (k - 1)
    sum_distr <- sum_distr + tdistr
    sum <- c(sum, sum_distr)
    k <- k + 1
  }

  sum
}
```

Затем, чтобы смоделировать случайную величину, распределённую по логарифмическому закону, передаём эту таблицу в обозначенную ниже функцию, которая моделирует равномерно распределённую величину и находит значение функции, обратной к функции распределения, от неё:

```
rlog <- function(sum){
  which.min(sum < runif(1))
}
```

Промоделированное логарифмическое распределение с параметром представлено на рис. А.1.

## А.2. Моделирование сложных распределений в R

Напомним, что сложное распределение — это случайная сумма случайных величин, поэтому для моделирования выборки из $n$ элементов этого распределения, нужно смоделировать выборку из $n$ элементов первого простого распределения, обозначающего количество слагаемых, для каждого из которых вычисляется сумма смоделированной

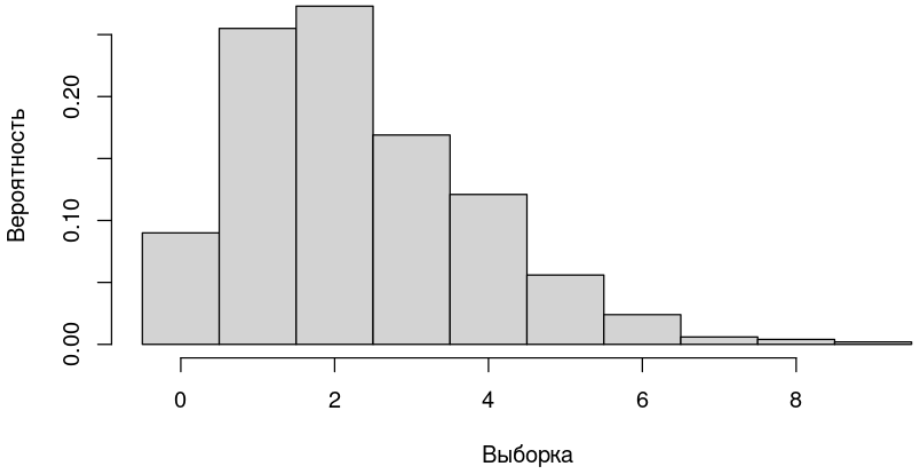


Рис. А.2. Промоделированное биномиально-логарифмическое распределение. $q = 0.2, p = 0.25$ и $n = 8$

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)))
        }

        res
}
```

Промоделированное биномиально-логарифмическое распределение с рассеянием меньше 1 представлено на рис. А.2.

## A.3. Оценка параметров и проверка согласия

Для оценки параметров применяется метод максимального правдоподобия, для чего, в первую очередь, составляется функция правдоподобия (пример приведём для биномиально-логарифмического распределения):

```
m_log_lik_binomlog <- function(x.in, p = 0.5, n = 1, q = 0.5){
        if (q <= 0 || q >= 1 || p <= 0 || p >= 1 || n <= 0){
                return(-log(0));
        }

        -sum(sapply(x.in, function(i)
         log(get_prob_binomlog(i, p, n, q))))
}
```

Часто приходится работать с частотами дискретных распределений, а не самой выборкой, поэтому для применения в функциях правдоподобия приведём функцию для создания выборки по частотам:

```
generate_sample <- function(num_k){
        rep.int(0:(length(num_k) - 1), num_k)
}
```

Тест $\chi^2$ по умолчанию не умеет объединять состояния для улучшения применимости, поэтому напишем свою статистику:

```
my_chisq <- function(exp_prob, prob){
        sum((exp_prob - prob)^2 / prob)
}
```

Теперь приведём основную функцию всей программы, которая производит оценку параметров, проверяет согласие и строит гистограмму:

```
hist_make <- function (n, exp_prob, get_hist, distr = 'binomlog'){
  sam <- generate_sample(exp_prob)
  N <- length(sam)
  var_sam <- var(sam) * (N - 1) / N
```

```
mean_sam <- mean(sam)
exp_prob <- exp_prob / N

if (n == 0){
  df <- -1
}
else{
  df <- 0
}

# Estimation of parameters for different distributions
# Binom-log
if (distr == 'binomlog'){
      # If the value of parameter n is set to zero,
      # then its evaluation is also performed
  if (n == 0){
    left <- 1
    right <- 1000

    # The ternary search method
    while (right - left > 2) {
      nLeft <- (2 * left + right) %/% 3
      nRight <- (left + 2 * right) %/% 3

      qLeft <- optim(c(0.5, 1 / nLeft), function(x)
       m_log_lik_binomlog(sam, p = x[2], n = nLeft, q = x[1]))
      qRight <- optim(c(0.5, 1 / nRight), function(x)
       m_log_lik_binomlog(sam, p = x[2], n = nRight, q = x[1]))

      if (qLeft$value <= qRight$value) {
        right <- nRight
```

```
      }
      else{
        left <- nLeft
      }
    }

    minValue <- optim(c(0.5, 1 / (left + 1)), function(x)
    m_log_lik_binomlog(sam, p = x[2], n = left, q = x[1]))
    q <- minValue$par[1]
    p <- minValue$par[2]
    n <- left

    for (i in (left + 1):right){
      value <- optim(c(0.5, 1 / i), function(x)
       m_log_lik_binomlog(sam, p = x[2], n = i, q = x[1]))

    if (value$value <= minValue$value){
      minValue <- value
      q <- minValue$par[1]
      p <- minValue$par[2]
      n <- i
    }
  }
}
# If n is given, then the score is only p and q
else {
  q <- optimize(function(x)
   m_log_lik_binomlog(sam, p = (var_sam / mean_sam *
   log(1 - x) * (1 - x) - log(1 - x)) / x, n = n, q = x),
   c(0.01, 0.99))$minimum
  p <- (var_sam / mean_sam * log(1 - q) * (1 - q) -
```

```
    log(1 - q)) / q
}

  res <- c(n, q, p)
}
# Log-binom
else if (distr == 'logbinom'){
  if (n == 0){
    left <- 1
    right <- 1000

    while (right - left > 2) {
      nLeft <- (2 * left + right) %/% 3
      nRight <- (left + 2 * right) %/% 3

      qLeft <- optim(c(0.5, 1 / nLeft), function(x)
       m_log_lik_logbinom(sam, q = x[1], p = x[2], n = nLeft))
      qRight <- optim(c(0.5, 1 / nRight), function(x)
       m_log_lik_logbinom(sam, q = x[1], p = x[2], n = nRight))

      if (qLeft$value <= qRight$value) {
        right <- nRight
      }
      else{
        left <- nLeft
      }
    }

    minValue <- optim(c(0.5, 1 / (left + 1)), function(x)
     m_log_lik_logbinom(sam, q = x[1], p = x[2], n = left))
    q <- minValue$par[1]
```

```
    p <- minValue$par[2]
    n <- left

    for (i in (left + 1):right){
      value <- optim(c(0.5, 1 / i), function(x)
       m_log_lik_logbinom(sam, q = x[1], p = x[2], n = i))

      if (value$value <= minValue$value){
        minValue <- value
        q <- minValue$par[1]
        p <- minValue$par[2]
        n <- i
      }
    }
  }
  else {
    q <- optimize(function(x)
     m_log_lik_logbinom(sam, q = x, p = - log(1 - x) *
     (1 - x) / n / x * mean_sam, n = n),
     c(0.01, 0.99))$minimum
    p <- -log(1 - q) * (1 - q) / n / q * mean_sam
  }

  res <- c(n, q, p)
# Log-pois
} else if (distr == 'logpois'){
  q <- optim(c(0.5, 1), function(x)
   m_log_lik_logpois(sam, q = x[1], lambda = x[2]))
  lambda <- q$par[2]
  q <- q$par[1]
  res <- c(lambda, q)
```

```
# Negative binom
} else if (distr == 'nbinom'){
  q <- optim(c(0.5, 1), function(x)
   m_log_lik_nbinom(sam, q = x[1], n = x[2]))
  n <- q$par[2]
  q <- q$par[1]
  res <- c(n, q)
# The sum of two negative binomials
} else if (distr == "doublenbinom"){
  q <- optim(c(0.8, 1, 0.5, 1),
   function(x) m_log_lik_doublenbinom(sam, p_1 = x[1],
   size_1 = x[2], p_2 = x[3], size_2 = x[4]))
  res <- c(q$par[1], q$par[2], q$par[3], q$par[4])
}

prob <- c()

# Getting accurate probabilities based on estimates
for (i in 0:(length(exp_prob) - 2)){
  if (distr == 'binomlog'){
    prob <- c(prob, get_prob_binomlog(i, p = res[3],
     n = res[1], q = res[2]))
  } else if (distr == 'logbinom'){
    prob <- c(prob, get_prob_logbinom(i, q = res[2],
     p = res[3], n = res[1]))
  } else if (distr == 'logpois'){
    prob <- c(prob, get_prob_logpois(i, q = res[2],
     lambda = res[1]))
  } else if (distr == 'nbinom'){
    prob <- c(prob, dnbinom(i, size = res[1],
     prob = res[2]))
```

```
  } else if (distr == 'doublenbinom'){
    prob <- c(prob, get_prob_doublenbinom(i, p_1 = res[1],
     size_1 = res[2], p_2 = res[3], size_2 = res[4]))
  }
}

prob <- c(prob, 1 - sum(prob))

max_el_x = max(sam)
max_el_y = max(exp_prob, prob)

# Histogram on demand
if (get_hist){
  hist.default(sam, probability = TRUE, breaks = seq(-0.5,
   max_el_x + 0.5, 1), xlim = c(-0.5, max_el_x + 0.5),
   ylim = c(0, max_el_y), xlab = "Sample", main = "")
  points(0:(length(prob) - 1), prob)
}

# Combining states
for (i in 1:length(prob)){
  while (100 * prob[i] < 5 && i < length(prob)){
    prob[i] <- prob[i] + prob[i + 1]
    prob <- prob[-(i + 1)]
    exp_prob[i] <- exp_prob[i] + exp_prob[i + 1]
    exp_prob <- exp_prob[-(i + 1)]
  }
}

if (100 * prob[length(prob)] < 5){
  prob[length(prob) - 1] <- prob[length(prob) - 1] +
```

```
    prob[length(prob)]
  prob <- prob[-length(prob)]
  exp_prob[length(exp_prob) - 1] <-
   exp_prob[length(exp_prob) - 1] + exp_prob[length(exp_prob)]
  exp_prob <- exp_prob[-length(exp_prob)]
 }

 # Counting degrees of freedom
 df <- df + length(prob) - 3

 if (df < 1){
  df <- 1
 }

 # Chi Square test
 chi <- N * my_chisq(exp_prob, prob)
 res <- c(res, 1 - pchisq(chi, df))

 res
}
```