Оценка параметров сложных распределений с применением в радиобиологии

Олейник Михаил Владимирович, гр.20.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к. ф.-м. н., доцент Алексеева Н. П.
Рецензент: научный сотрудник, СПбГУ, ФМКН, Белоусов Ю. С.

Санкт-Петербург, 2024

Введение

Определение

Сложное распределение — случайная сумма независимых одинаково распределенных случайных величин.

Основное направление исследования сложных распределений — описание ветвящихся процессов. Применение в физике частиц. Другие возможные области применения сложных распределений:

1. Радиобиология (исследование ядерных аномалий, работа [Алексеева, 2008]);
2. Анализ текстов (встречаемость слов в текстах, работа [Alexeeva, Sotov, 2013])

Цель работы: оценка параметров и проверка согласованности различных сложных распределений с эмпирическими данными в радиобиологии и анализе текстов.

Производящая функция и рассеяние

Производящая функция дискретного распределения:

$$h(t) = p_0 + p_1 t + p_2 t^2 + \ldots, \quad P(\xi = j) = p_j, \quad j = 0, 1, 2, \ldots.$$

Свойства [Феллер, 1952]:

1. $\mathsf{E}\xi = h'(1)$, $\mathsf{D}\xi = h''(1) + h'(1) - (h'(1))^2$;
2. $P(\xi = k) = \frac{h^{(k)}(0)}{k!}$.

Лемма

S_τ — сумма случайного числа τ случайных величин ξ_i, одинаково распределённых и независимых между собой и τ. Для её рассеяния ($e(S_\tau) = \mathsf{D}S_\tau / \mathsf{E}S_\tau$) справедлива формула:

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

Если $e(\xi_i) \geqslant 1$, то $e(S_\tau) \geqslant 1$ $(P(\xi_i \geqslant 0) = 1)$.

Поэтому ищем сложные распределения с переменным знаком логарифма рассеяния.

Рассматриваемые сложные распределения

Вид сложных распределений в работе:

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

где ξ_i одинаково распределены и независимы между собой и τ. При разных вариациях τ и ξ_i рассмотрены следующие распределения:

1. ζ_1 — Биномиально(τ)-логарифмическое(ξ_i) (БЛР);
2. ζ_2 — Логарифмически(ξ_i)-биномиальное(τ) (ЛБР);
3. η — Логарифмически(ξ_i)-пуассоновское(τ) (ЛПР).

Числа Стирлинга

- **Первого рода** — количество перестановок длины k с j циклами.

Рекуррентная формула:

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

$k \setminus j$	0	1	2	3	4
0	1				
1	0	1			
2	0	1	1		
3	0	2	3	1	
4	0	6	11	6	1

- **Второго рода** — количество неупорядоченных разбиений k-элементного множества на j непустых подмножеств.

Явная формула:

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

Рекуррентная формула:

$$S(k+1,j) = S(k,j-1) + j \cdot S(k,j);$$
$$S(0,0) = 1, S(k,0) = 0, S(k,j) = 0$$

$k \setminus j$	0	1	2	3	4
0	1				
1	0	1			
2	0	1	1		
3	0	1	3	1	
4	0	1	7	6	1

Числа Эйлера

Числа Эйлера первого рода — количество перестановок длины k с j подъемами.

Явная формула: $E(k,j) = \sum\limits_{i=0}^{j} C_{k+1}^{i}(-1)^i(j+1-i)^k$. Рекуррентная формула:

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

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

$k \setminus j$	0	1	2	3	4
0	1				
1	1				
2	1	1			
3	1	4	1		
4	1	11	11	1	
5	1	26	66	26	1

Возникают в рядах: $\sum\limits_{i=1}^{\infty} i^k x^i = \frac{x}{(1-x)^{k+1}} \sum\limits_{j=0}^{k-1} E(k,j)x^j$.

Биномиально-логарифмическое распределение

Введём новое сложное распределение:

- $\zeta_1 = \xi_1 + \ldots + \xi_\tau$.
- $\xi_i \sim \mathbf{Log}(\cdot, q)$ — независимые случайные величины.

$$P\{\xi_i = j\} = \frac{\alpha q^j}{j}, \quad \text{где } \alpha = -\big(\ln(1-q)\big)^{-1}, \quad j = 1, 2, \ldots.$$

- $\tau \sim \mathbf{Bin}(\cdot|n, p)$ — случайная величина, не зависящая от ξ_i
- Тогда $\zeta_1 \sim \mathbf{BinLog}(\cdot|n, p, q)$.
- Производящая функция: $h_1(t) = \big((1-p) - \alpha p \ln(1-qt)\big)^n$.
- При $n \to \infty, p \to 0$ — негативный бином (обобщение).

Теорема

$$P(\zeta_1 = k) = \frac{1}{k!}(1-p)^{n-k} \cdot q^k \sum_{j=0}^{k} \frac{n!}{(n-j)!} s(k, j)(p\alpha)^j (1-p)^{k-j}.$$

Логарифмически-биномиальное распределение

Обратная ситуация:

- $\zeta_2 = \xi_1 + \ldots + \xi_\tau$.
- $\xi_i \sim \mathbf{Bin}(\cdot|n,p)$ — независимые случайные величины.
- $\tau \sim \mathbf{Log}(\cdot,q)$ — случайная величина, не зависящая от ξ_i
- Тогда $\zeta_2 \sim \mathbf{LogBin}(\cdot|n,p,q)$.
- Производящая функция:

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

- Вероятности вычислены до $k = 4$:

$$P(\zeta_2 = k) = \frac{1}{k!} h_2^{(k)}(0).$$

Модель логарифмически-пуассоновского распределения

Если в ЛБР устремить n к бесконечности, а p к нулю, то получим новое распределение — логарифмически-пуассоновское с параметрами q и $\lambda = p \cdot n$.

Определение

- $\xi_1 + \ldots + \xi_\tau = \eta \ \sim \ \mathbf{LogPois}(\cdot|\lambda, q)$, *если*
- $\xi_i \sim \mathbf{Pois}(\cdot|\lambda)$ — *независимые случайные величины,*
- $\tau \sim \mathbf{Log}(\cdot|q)$ — *случайная величина, не зависящая от* ξ_i.
- *Производящая функция сл.в.* η

$$h(t) = -\alpha \ln(1 - qe^{\lambda(t-1)}), \quad \alpha = -\big(\ln(1-q)\big)^{-1}.$$

Вероятности ЛПР

Теорема 1

Вероятности ЛПР:

$$P(\eta = 0) = -\alpha \log\left(1 - qe^{-\lambda}\right),$$

$$P(\eta = k) = \frac{\alpha}{k!}\lambda^k \sum_{j=1}^{k}(j-1)!S(k,j)\left(\frac{qe^{-\lambda}}{1-qe^{-\lambda}}\right)^j, \text{ при } k = 1, 2\ldots.$$

Теорема 2

Для $k = 0, 1$ формула аналогична теореме 1.

$$P(\eta = k) = \frac{\alpha}{k!}\frac{\lambda^k qe^{-\lambda}}{\left(1-qe^{-\lambda}\right)^k}\sum_{j=0}^{k-2}E(k-1,j)\left(qe^{-\lambda}\right)^j, \quad k > 1.$$

Для реальных вычислений используются рекуррентные формулы и нормированные числа Стирлинга и Эйлера.

Тестовые данные

Массив данных о количестве аномалий в ядрах клеток у крыс при различных дозах облучения in vivo (на живую) и in vitro (облученные клетки подсажены здоровым крысам).

- В [Алексеева, 2008] в качестве модели для количества аномалий предложено реинтрантно-биномиальное распределение, которое имело четыре параметра, что много, учитывая оценку и интерпретацию параметров по центральным моментам.
- Задача апробации моделей с различным количеством параметров: БЛР, ЛБР (3 параметра) и ЛПР (2 параметра).

Оценки in vivo, in vitro: БЛР

Гр	n	q	p	p-v
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: in vivo

Гр	n	q	p	p-v
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

Таблица: in vitro

- У найденных оценок не наблюдается монотонной зависимости от дозы облучения.
- Для подавляющего большинства БЛР (84%) даёт согласованность на уровне $\alpha = 0.05$.

Оценки in vivo, in vitro: ЛБР

Гр	n	q	p	p-v
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	999	0.32	1.5e-3	0.31
35	998	0.19	1.7e-3	0.11
40	999	0.22	2.1e-3	0.16
45	998	0.18	2.3e-3	0.08

Таблица: in vivo

Гр	n	q	p	p-v
0	20	1.7e-6	0.02	0.52
5	994	0.80	1.4e-4	0.73
10	999	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

Таблица: in vitro

- У найденных оценок не наблюдается монотонной зависимости от дозы облучения.
- Для подавляющего большинства ЛБР (89%) даёт согласованность на уровне $\alpha = 0.05$.

Оценки in vivo, in vitro: ЛПР

Для радиобиологических данных о числе аномалий на ядрах рабдомиосаркомы при разной степени облучения получены ОМП модели ЛПР и проверено согласие по χ^2.

Таблица: In vivo

Гр	λ	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

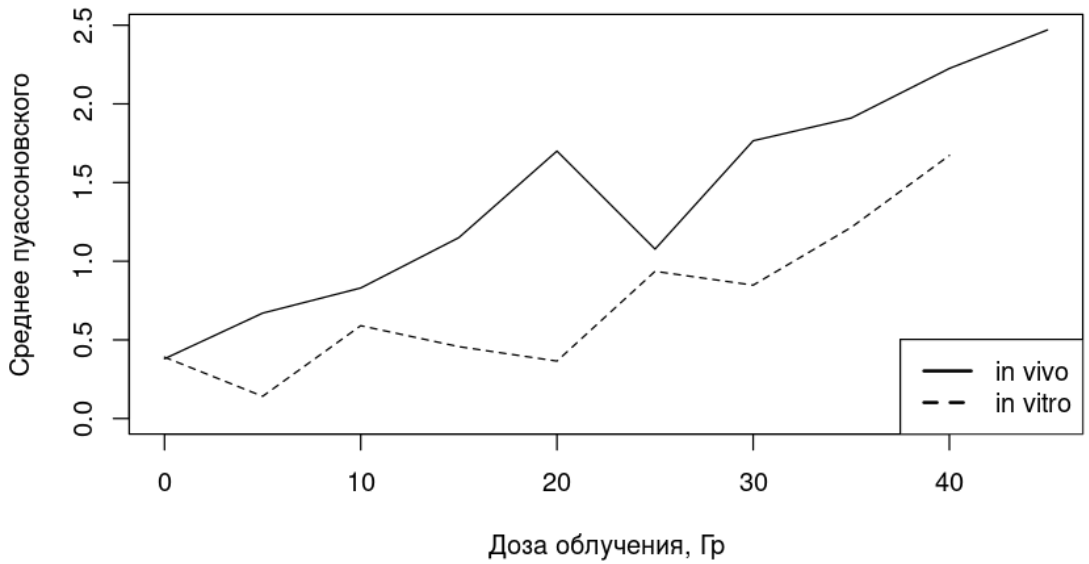
Таблица: In vitro

Гр	λ	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

При уровне значимости $\alpha = 0.05$ получаем согласие для $16/19 \cdot 100\% = 84\%$ случаев.

Интерпретация параметра λ распределения Пуассона

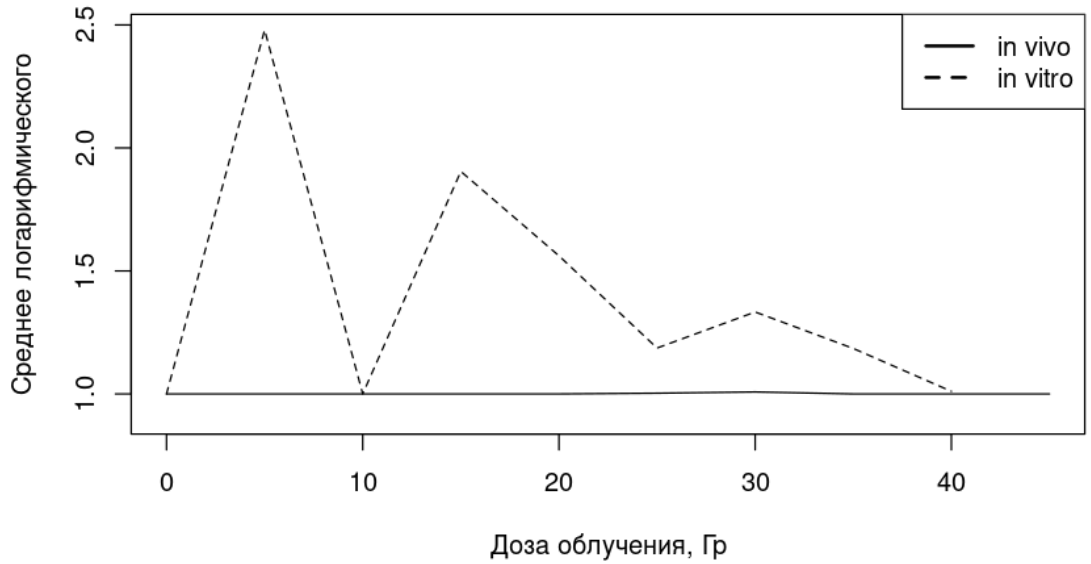
- Количество аномалий определяется двумя факторами: **исходной распространенностью аномалий** (параметр q) и **интенсивностью образования** (λ).
- Динамика средних значений Пуассона свидетельствует о **положительной** линейной зависимости от дозы и о меньших значениях in vitro, так как выжившие клетки обладают **большим иммунитетом**.



Интерпретация параметра q логарифмического распределения

- Чем больше q, тем выше распространенность аномалий.
- В in vitro распространенность значимее, чем в in vivo.
- От дозы облучения зависит **количество выживших клеток**, а **не распространенность** их аномалий.

При небольших дозах облучения в in vitro распределения аномалий носят **экстенсивный** характер, а при очень высоких **интенсивный**.



Встречаемость слов

Вторая задача работы — анализ текстов.

- Задача: дан текст из n глав. Некоторое слово встречается в i-ой главе x_i раз.
- Вопрос: какому распределению удовлетворяет выборка $(x_1, \ldots, x_n)$?
- Ответ: неплохое согласование даёт модель отрицательного бинома [Alexeeva, Sotov, 2013].
- Однако есть ряд слов, не согласующихся с ОБР, поэтому возникает идея проверить их согласованность с БЛР, как обобщением ОБР.
- Проблемные слова имеют распределения с тяжёлыми хвостами, БЛР и ОБР дают плохую согласованность;
- Предположение — употребление слов в нескольких значениях: более частых или менее;
- Рассмотрим сумму двух независимых ОБР величин:

$$\eta = \xi_1 + \xi_2, \quad \xi_1, \xi_2 \sim NB(r, q) \text{ и нез.}$$

Классы слов

- Для проверки согласованности взят роман Теодора Драйзера «Американская трагедия» на английском языке;
- Размер выборки (количество глав): $n = 102$;
- Проанализировано первые по встречаемости 1000 слов.

При уровне значимости $\alpha = 0.05$ разобьём все слова на **не пересекающиеся классы** по согласованности с распределениями:

Распределение	Согласованные слова
ОБР	4
БЛР	7
Сумма	14
ОБР и БЛР	22
ОБР и сумма	49
БЛР и сумма	10
Все	839
Итого	945

Результаты

- **Малая часть** (0.031) слов удовлетворяют сумме ОБР или БЛР, но не ОБР. Эти распределения **не сильно** расширяют класс применимых моделей;
- Общая согласованность составляет 0.945, что почти равняется $\gamma = 1 - \alpha = 0.95$;
- Слова располагаются на **двумерной поверхности**, судя по точечному графику. Частично это обусловлено средним встречаемости слов.

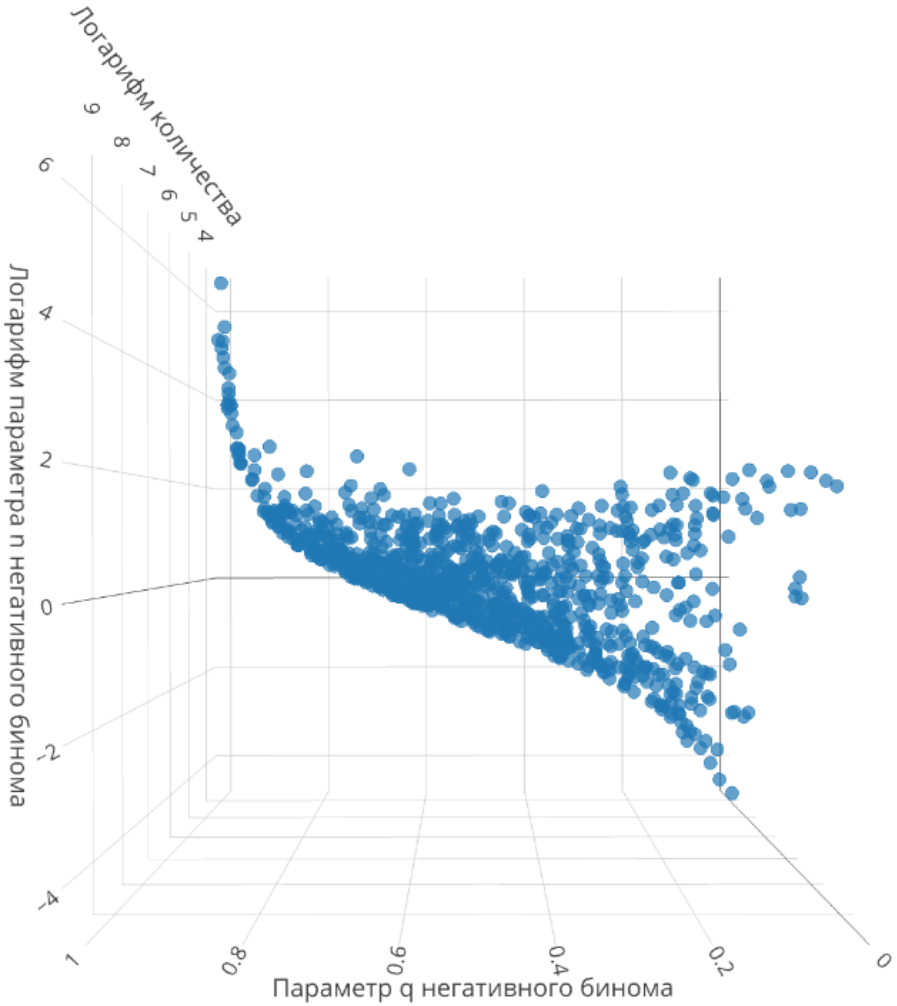


Рис.: Точечный график слов.

Заключение

- Были рассмотрены трёхпараметрические БЛР и ЛБР и двухпараметрическое ЛПР, показана их применимость к эмпирическим данным из работы [Алексеева, 2008].
- Лучшей моделью можно считать ЛПР, так как она имеет наименьшее количество параметров и примерно такое же согласование, что и остальные.
- Проверено 3 модели к встречаемости слов в тексте. ОБР даёт согласованность с 91% слов, БЛР и сумма ОБР расширяют множество согласованных слов на 3.1%. Доля согласованных слов близка к теоретическому значению.
- Для дальнейшего исследования интересны вопросы:
  1. Границы применимости распределений с рассеянием большим 1 (ЛПР) к данным с рассеянием меньше 1 (ряд случаев в радиобиологии).
  2. Роль различных классов чисел в формулах вероятностей сложных распределений.

Возникновение классов чисел

τ распределено по $(P_0, P_1, P_2, \ldots)$, а ξ_i — по $(p_0, p_1, p_2, \ldots)$:

$$P\{\zeta_\tau = n\} = \frac{1}{n!}\sum_{k=1}^{n}\Theta_k G_n^k\,, \quad \text{где} \quad \Theta_k = \sum_{i=k}^{\infty} P_i C_i^{i-k} p_0^{i-k}\,,$$

$$G_n^k = n! \sum_{\sum_i^k n_i = n} \prod_{i=1}^{k} p_{n_k}$$

$$\textbf{Log} : u_n^k = \sum_{\sum_i^k n_i = n, n_i > 0} \frac{n!}{n_1 \ldots n_k}, \quad \textbf{Pois} : v_n^k = \sum_{\sum_i^k n_i = n, n_i > 0} \frac{n!}{n_1! \ldots n_k!}$$

с рекуррентными соотношениями

$$u_n^k = k u_{n-1}^{k-1} + (n-1) u_{n-1}^k\,, \qquad v_n^k = k(v_{n-1}^{k-1} + v_{n-1}^k)\,,$$

которые приводят к существенным частным случаям в виде чисел Стирлинга первого и второго рода $s(n,k) = \frac{1}{k!}u_n^k$ и $S(n,k) = \frac{1}{k!}v_n^k$.