Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского»

Кафедра Математических Методов Системного Анализа

Расчетная работа с дисциплины "Математическая статистика"

Проверил
к.ф.м.н. Каниовская И. Ю.
Выполнил
студент и просто хороший человек Панченко Е. С.

Киев 2021

Содержание

1. Постановка задачи

Задана выборка

2	1	3	1	0	1	3	2	2	2
4	5	1	3	2	2	2	4	2	2
4	5	6	4	3	5	2	4	2	4
3	3	5	9	6	3	5	7	1	6
3	2	1	4	9	5	4	5	4	5
4	4	9	2	1	6	1	5	3	5
5	3	6	9	3	6	4	4	4	8
3	7	4	5	2	6	4	3	3	3
4	2	4	2	3	5	4	3	4	5
2	2	6	3	5	2	6	2	1	1

Основная цель - найти, каким распределением она порождена и аргументировать, почему.

2. Решение задачи

2.1. Построение вариационного ряда данной выборки

Поскольку в выборке содержатся только целые числа и большинство из них не уникально, то автор решил построить дискретный ряд. Для этого строится табличка, которая по каждой варианте показывет ее частость и частотность.

варианты	0	1	2	3	4	5	6	7	8	9
частоты n_i	1	10	20	18	20	15	9	2	1	4
частотности w_i	0.01	0.1	0.2	0.18	0.2	0.15	0.09	0.02	0.01	0.04

2.2. Графическое изображение выборки

Изображение выборки было решено представить в виде гистограммы, где каждый столбец представляет собой варианту, а высота каждого столбца - ее частость. Гистограмма имеет вид:

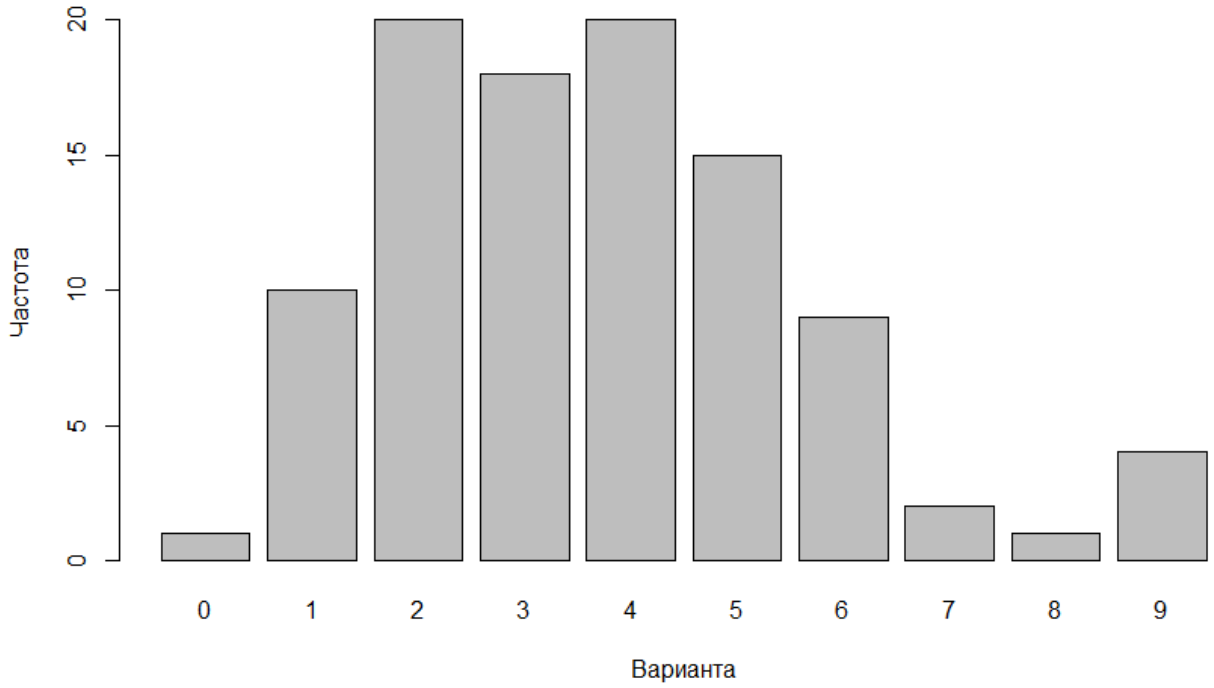

Рис. 1. Гистограмма выборки

2.3. Эмпирическая функция распределения

Построим эмпирическую функцию распределения:

$$F^*_{100}(x) = \begin{cases} 0, & \text{если } x \le 0; \\ 0.01, & \text{если } 0 < x \le 1; \\ 0.01 + 0.1 = 0.11, & \text{если } 1 < x \le 2; \\ 0.01 + 0.1 + 0.2 = 0.31, & \text{если } 2 < x \le 3; \\ 0.01 + 0.1 + 0.2 + 0.18 = 0.49, & \text{если } 3 < x \le 4; \\ 0.01 + 0.1 + 0.2 + 0.18 + 0.2 = 0.69, & \text{если } 4 < x \le 5; \\ 0.01 + 0.1 + 0.2 + 0.18 + 0.2 + & \\ +0.15 = 0.84, & \text{если } 5 < x \le 6; \\ 0.01 + 0.1 + 0.2 + 0.18 + 0.2 + & \\ +0.15 + 0.09 = 0.93, & \text{если } 6 < x \le 7; \\ 0.01 + 0.1 + 0.2 + 0.18 + 0.2 + & \\ +0.15 + 0.09 + 0.02 = 0.95, & \text{если } 7 < x \le 8; \\ 0.01 + 0.1 + 0.2 + 0.18 + 0.2 + & \\ +0.15 + 0.09 + 0.02 + 0.01 = 0.96, & \text{если } 8 < x \le 9; \\ 0.01 + 0.1 + 0.2 + 0.18 + 0.2 + & \\ +0.15 + 0.09 + 0.02 + 0.01 + 0.04 = 1, & \text{если } x > 9. \end{cases}$$

Убрав промежуточные расчеты, получим

$$F^*_{100}(x) = \begin{cases} 0, & \text{если } x \le 0; \\ 0.01, & \text{если } 0 < x \le 1; \\ 0.11, & \text{если } 1 < x \le 2; \\ 0.31, & \text{если } 2 < x \le 3; \\ 0.49, & \text{если } 3 < x \le 4; \\ 0.69, & \text{если } 4 < x \le 5; \\ 0.84, & \text{если } 5 < x \le 6; \\ 0.93, & \text{если } 6 < x \le 7; \\ 0.95, & \text{если } 7 < x \le 8; \\ 0.96, & \text{если } 8 < x \le 9; \\ 1, & \text{если } x > 9. \end{cases}$$

Ниже приводится график эмпирической функции распределения.

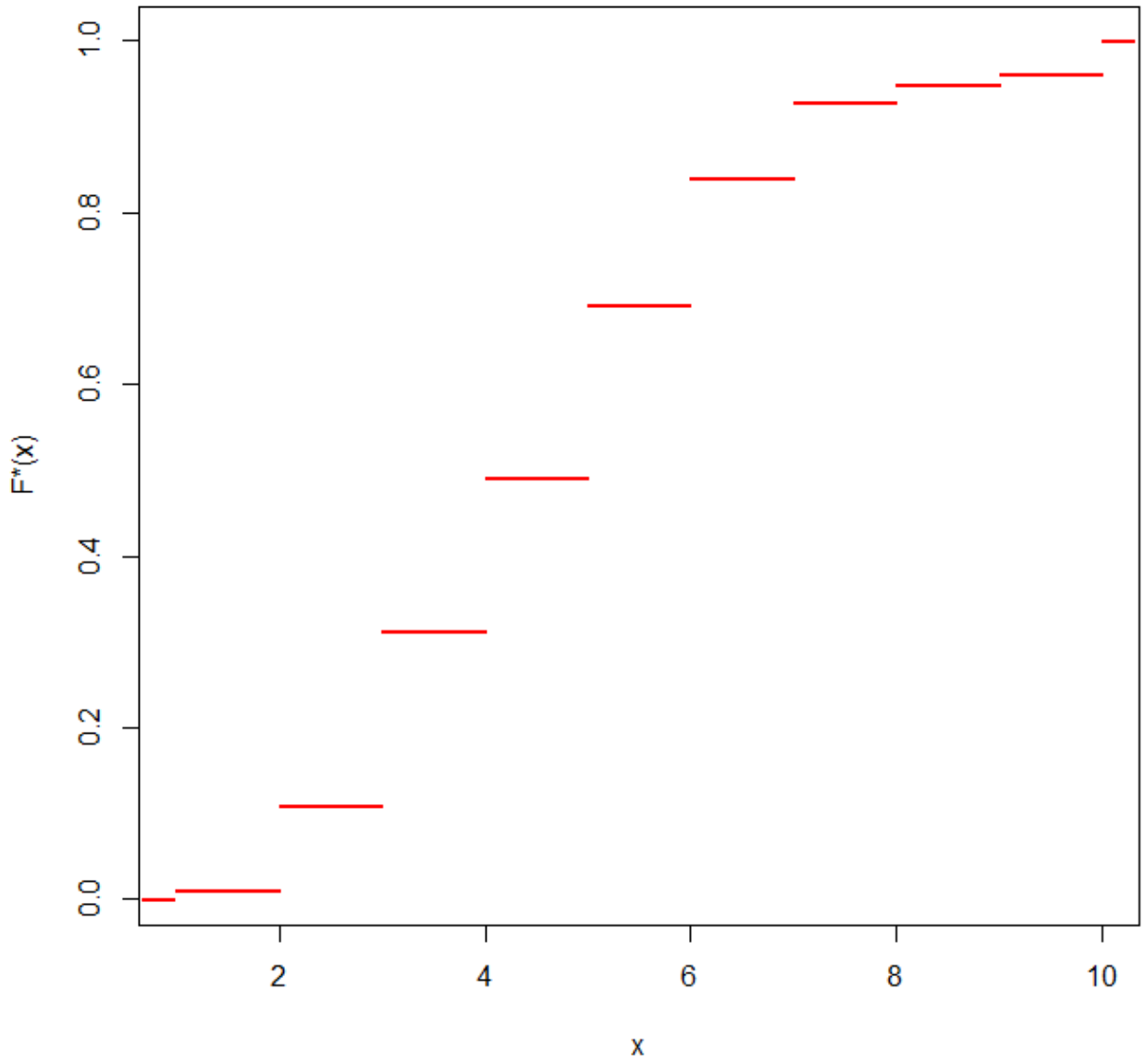

Рис. 2. График эмпирической функции распределения

2.4. Выборочные медиана, мода и ассиметрия

Выборочная медиана есть медиана выборки. То есть для ее вычисления достаточно взять среднее арифметическое 50-го и 51-го элементов, когда выборка предварительно упорядочена:

$$Me^*(\xi) = \frac{4+4}{2} = 4.$$

Выборочной модой есть те варианты, которые в выборке встречаются наибольшее число раз. В данной выборке есть две варианты, которые встречаются

чаще других - это 2 и 4. То есть

$$Mo^*(\xi) = \{2, 4\}.$$

Выборочная ассиметрия считается по формуле

$$As^*(\xi) = \frac{\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^3}{(\mathbb{D}^*\xi)^{3/2}},$$

где $\overline{\xi}$ - выборочное среднее и $\mathbb{D}^*\xi$ - выборочная диспресия. Считаем значения выборного среднего и выборочной дисперсии на данной выборке:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{100} x_i = 3.71,$$

$$(\mathbb{D}^*\xi)_{val} = \frac{1}{n}\sum_{i=1}^{100}(x_i - \overline{x})^2 = 3.844343,$$

$$(As^*\xi)_{val} = \frac{\frac{1}{100}\sum_{i=1}^{100}(x_i - 3.71)^3}{(3.844343)^{3/2}} \approxeq 0.709.$$

2.5. Нахождение несмещенных оценок математического ожидания и дисперсии

Известно, что исправленная выборочная дисперсия - это несмещенная оценка дисперсии генеральной совокупности. В дальнейшей работе этого факта нам будет вполне достаточно. Подробно про несмещенную оценку математического ожидания будет написано в седьмом подзаголовке.

2.6. Выдвижение гипотезы про распределение, за которым получено выборку

Моё предположение, что выборка порождена распределением Пуассона. Для этого есть несколько причин. Самое главное замечание, что в выборке содержаться только целые числа. Из этого я делаю вывод, что распределение, скорее всего, дискретное. Геометрическое вряд-ли, потому что столбцы в гистограмме не убывают. Есть вариант, что это может быть и биномиальное распределение. Сначала я проверю, порождена ли эта выборка распределением Пуассона, и если нет, то перейду на биномиальное рапспределение. Также я поигрался с разными значениями параметров для распределения Пуассона и биномиального распределения. Ниже можно увидеть ряд графиков, которые подтверждают обе мои гипотезы.

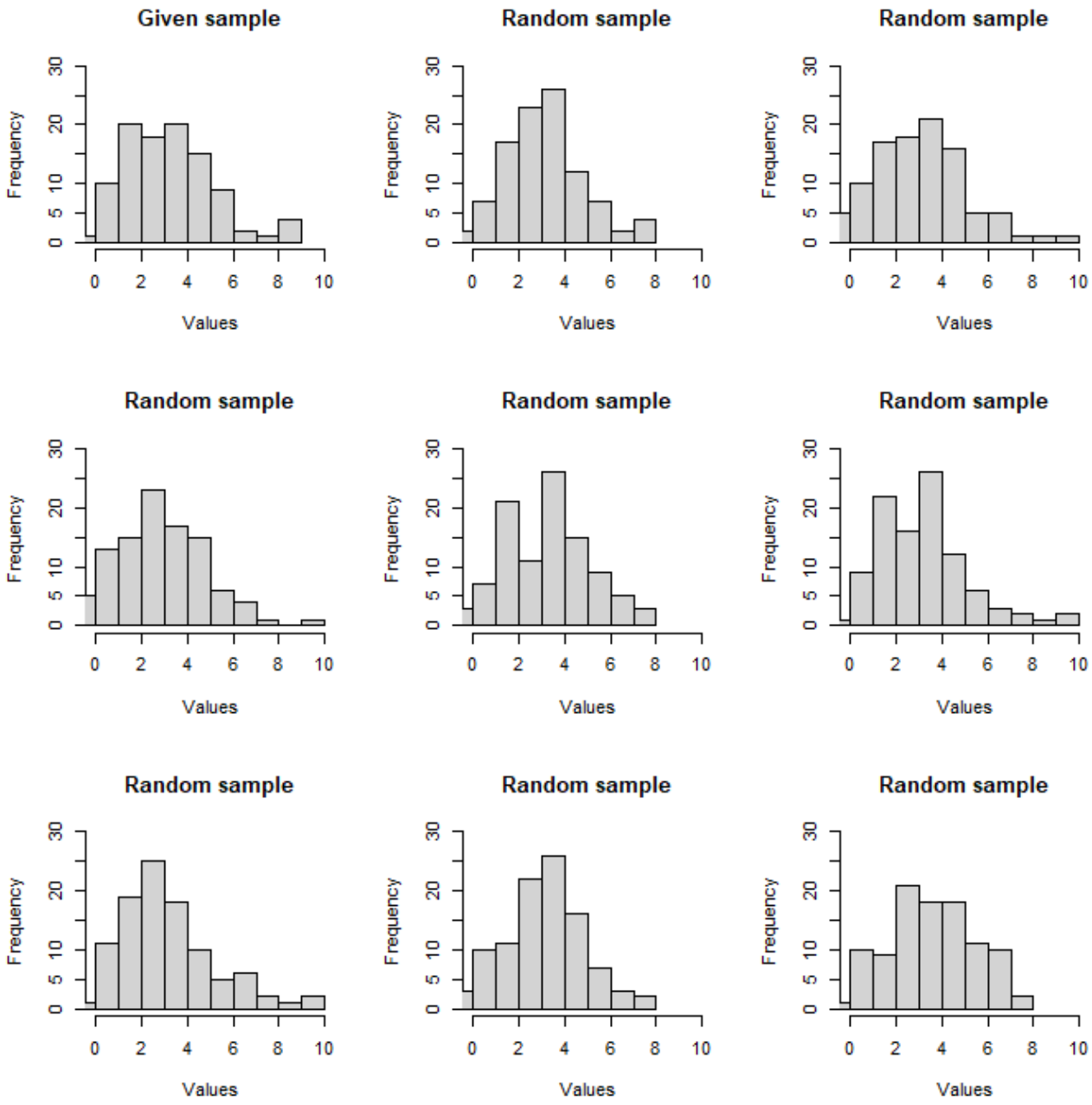

Рис. 3. Случайно сгенерированные выборки, попрожденные законом Пуассона с параметром $a = 3.71$

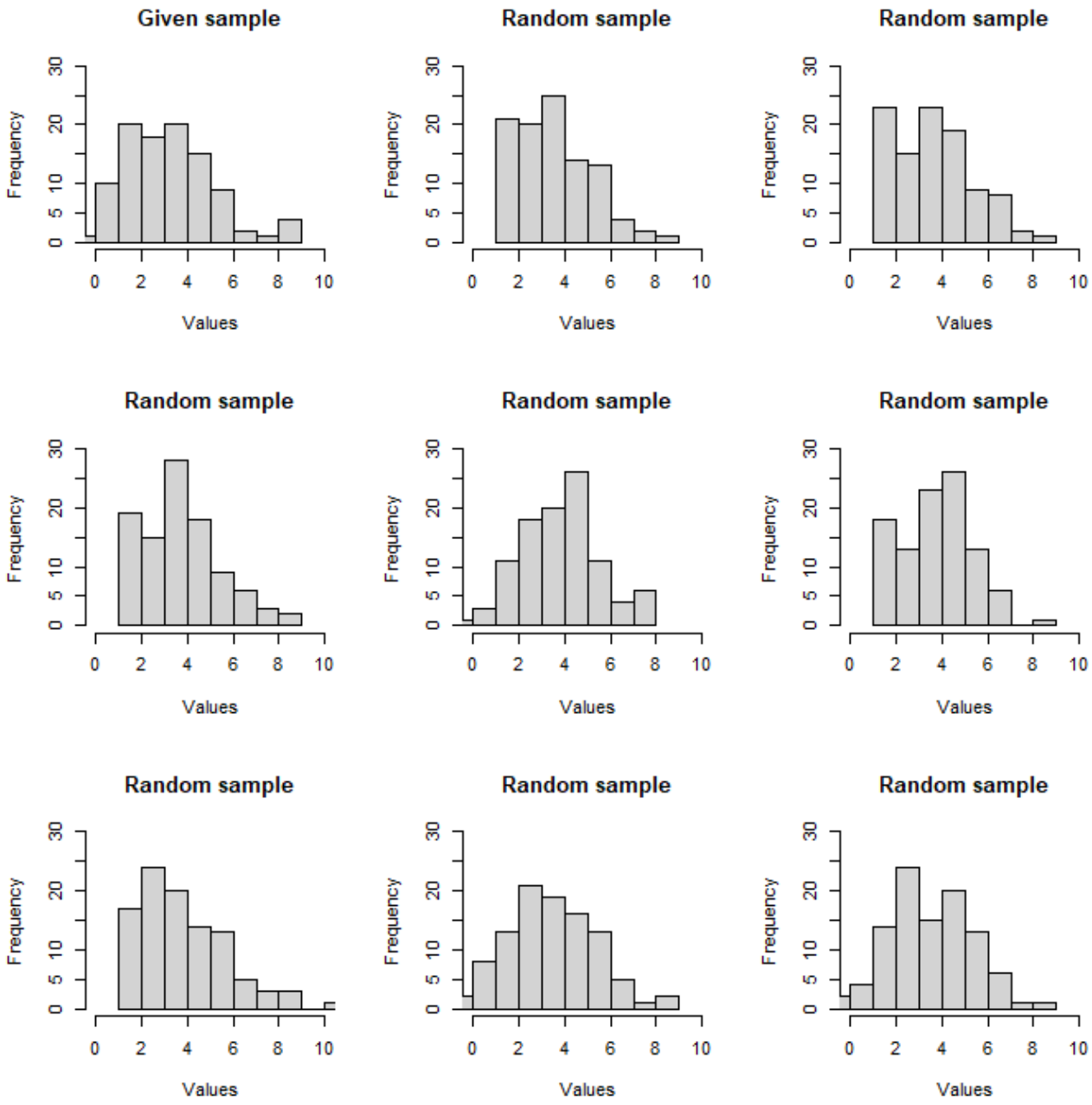

Рис. 4. Случайно сгенерированные выборки, порожденные биномиальным законом с параметрами $size = 20, p = 0.2$

2.7. Нахождение точковых оценок параметров распределения Пуассона и проверить их свойства

Пусть наша выборка порождена распределением Пуассона, то есть

$$\xi \sim Poiss(a).$$

Известно, что математическое ожидание случайной величины, распределенной по распределению Пуассона с параметром a, есть само число a. Исходя

из этого, выберем в качестве точечной оценки параметра a выборочное среднее. Исследуем свойства этой точечной оценки. Известно, что выборочное среднее есть несмещенной оценкой математического ожидания, а также эта оценка есть конзистентной. Проверим ее эффективность:

$$ln\mathcal{L}(\vec{x}, a) = -n \cdot a + ln(a) \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} ln(x_i!),$$
$$\frac{\partial ln\mathcal{L}(\vec{x}, a)}{\partial a} = -n + \frac{1}{a} \cdot \sum_{i=1}^{n} x_i = \frac{n}{a} \cdot (\overline{x} - a).$$

Поскольку множитель перед $(\overline{x} - a)$ зависит только от n и от a, то оценка является эффективной.

Подытожим: на данный момент у нас есть несмещенная, конзистентная и эффективная оценка параметра гипотетического закона распределения. Это уже неплохой старт. Найдем теперь точечную оценку параметра a двумя методами - методов моментов и методом максимальной правдоподобности.

2.7.1. Метод моментов

Суть метода моментов в том, что эмпирические моменты приравниваются к теоретическим моментам. Если из этих равенств можно вытянуть числовое значение некой характеристики, то это значение можно брать за значение точечной оценки этой характеристики. К делу, приравняем моменты первых порядков:

$$a^* = E\xi = E^*\xi = \overline{\xi}.$$

Получили, что $a^*_{\text{ММ}} = \overline{x} = 3.71$.

2.7.2. Метод максимальной правдоподобности

Суть метода максимальной правдоподобности в том, чтобы максимизировать значение функции правдоподобности. Функция правдоподобности распределения Пуассона есть

$$ln\mathcal{L}(\vec{x}, a) = -n \cdot a + ln(a) \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} ln(x_i!).$$

Найдем критические точки:

$$\frac{\partial ln\mathcal{L}(\vec{x},a)}{\partial a} = -n + \frac{1}{a}\cdot\sum_{i=1}^{n} x_i = \frac{n}{a}\cdot(\overline{x}-a) = 0 \iff a_{cr} = \overline{x}.$$

Чтобы проверить, что в точке a_{cr} достигается максимум, устремим в функции правдоподобности аргумент a к плюс бесконечности и к нулю. Тогда функция правдоподобности будет стремится к минус бесконечности в обоих случаях. А это значит, что в точке a_{cr} достигается максимум.

Получили, что $a^*_{\text{ММП}} = \overline{x} = 3.71$.

2.8. Проверка гипотезы про распределение с помощью критерия Пирсона

Пусть некая случайная величина распределена по закону Пуассона. Построим табличку, в которой будут хранится значения вероятностей, что эта случайная величина набыла одно из значений от нуля до девяти. Округлим вероятности до четырех знаков после запятой.

значение случ. вел.	0	1	2	3	4
вероятность	0.0244	0.0908	0.1684	0.2083	0.1932
значение случ. вел.	5	6	7	8	9
вероятность	0.1433	0.0886	0.0470	0.0218	0.0089

Поскольку из-за округления сумма вероятностей не равна единице, то подгоним значения вероятностей таким образом, чтобы их сумма стала единицей. Для этого сначала подсчитаем сумму всех этих вероятностей:

$$sum_{prob} = 0.9947 \text{ , а заодно и } \frac{1}{sum_{prob}} \approx 1.005.$$

Затем умножим каждое из чисел на на величину, обратную найденной суммы. Получим табличку

значение случ. вел.	0	1	2	3	4
вероятность	0.0247	0.0912	0.1692	0.2093	0.1943
значение случ. вел.	5	6	7	8	9
вероятность	0.1441	0.0891	0.0472	0.0219	0.009

Теперь умножим каждую из вероятностей на размер выборки, то есть на сотню.

значение случ. вел.	0	1	2	3	4
$n \cdot p$	2.47	9.12	16.92	20.93	19.43
значение случ. вел.	5	6	7	8	9
$n \cdot p$	14.41	8.91	4.72	2.19	0.9

Объясним, что сейчас будет происходить. Мы собираемся воспользоваться критерием Пирсона. Для этого мы вводим нулевую гипотезу H_0 - выборка порождена распределением Пуассона с параметром 3.71. Наша цель показать, что гипотеза H_0 выполняется. В таком случае роботу можно оканчивать, и не проверять биномиальное распределение.

По алгоритму проверки гипотезы по критерию Пирсона, необходимо разбить множество возможных значений случайной величины на несколько непересекающихся классов. Поскольку случайная величина набывает десяти значений, то начальное число классов равно десяти. Наша цель - объеденить некоторые классы так, чтобы в каждом из классов величина $n \cdot p$ была не меньше десяти. Для этого объеденим первые два класса и последние четыре. Перепишем нашу табличку, а также внесем в нее дополнительные строки - частоты на множествах значений случайной величины, и разницы соответствующих значений.

значение случ. вел. ξ_i	$\{0,1\}$	2	3	4	5	$\{6,7,8,9\}$
$n \cdot p_i$	11.59	16.92	20.93	19.43	14.41	16.72
частота n_i	11	20	18	20	15	16
разница $n_i - n \cdot p_i$	-0.59	1.08	-0.93	0.57	0.59	-0.72

Теперь необходимо подсчитать η_{val}. По определению

$$\eta = \sum_{i=1}^{r} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}.$$

Подставляя числа из таблицы, получим

$$\eta_{val} = \frac{(-0.59)^2}{11.59} + \frac{(1.08)^2}{16.92} + \frac{(-0.93)^2}{20.93} + \frac{(0.57)^2}{19.43} + \frac{(0.59)^2}{14.41} + \frac{(-0.72)^2}{16.72} \approxeq 0.2121.$$

По теореме Пирсона статистика η стремится по распределению к распределению χ^2_{r-s-1}, где r - число классов, а s - число неизвестных параметров. В нешем случае $r = 6$ и $s = 1$. Смотрим в таблицу квантилей распределения Пирсона, и находим, что $t_{cr} = 9.49$.

Поскольку $\eta_{val} < t_{cr}$, то можно порадоваться, ведь наши данные не противоречат выдвинутой гипотезе H_0.

2.9. Нахождения доверительного интервала гипотетического закона распределения

По центральной граничной теореме, приблизительно имеем, что

$$\overline{\xi} \sim N(E\xi, \frac{D\xi}{n}).$$

Или

$$\overline{\xi} \sim N(a, \frac{a}{n}).$$

Или, нормируя случайную величину:

$$\eta = \frac{(\overline{\xi} - a) \cdot \sqrt{n}}{\sqrt{a}} \sim N(0, 1).$$

Задача нахождения доверительного интервала состоит в том, чтобы отыскать такие числа t_1 и t_2, чтобы $P\{t_1 < a < t_2\}$ была не менее γ. По условию $\gamma = 0.95$. Для упрощения задачи будем искать именно симметричный интервал относительно точки a. Тогда наша вероятность перепишется в виде

$$P\{|\eta| < t_\gamma\} = \gamma,$$

где t_γ находится из таблицы Лапласса. Поскольку

$$P\{|\eta| < \varepsilon\} = 2 \cdot \Phi(\varepsilon) = \gamma = 0.95, \text{ откуда } \varepsilon = t_\gamma = 1.96.$$

Запишем цепочку преобразований

$$|\eta| < 1.96 \iff \frac{|\overline{\xi} - a| \cdot \sqrt{n}}{\sqrt{a}} < 1.96 \iff$$
$$\frac{|3.71 - a| \cdot 10}{\sqrt{a}} < 1.96 \iff (3.71 - a)^2 \cdot 100 < 1.96^2 \cdot a \iff$$
$$a^2 - 7.42 \cdot a + 13.7641 < 0.038416 \cdot a \iff 3.3512 < a < 4.10722.$$

То есть доверительный интервал есть $(3.3512, 4.10722)$.

2.10. Выводы