

**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

<b>1. Постановка задачи</b>	<b>3</b>
<b>2. Решение задачи</b>	<b>3</b>
2.1. Построение вариационного ряда данной выборки . . . . .	3
2.2. Графическое изображение выборки . . . . .	3
2.3. Эмпирическая функция распределения . . . . .	5
2.4. Выборочные медиана, мода и асимметрия . . . . .	6
2.5. Нахождение несмещенных оценок математического ожидания и дисперсии . . . . .	7
2.6. Выдвижение гипотезы про распределение, за которым получено выборку . . . . .	7
2.7. Нахождение точечной оценки параметра распределения Пуассона и проверка его свойств . . . . .	9
2.7.1. Метод моментов . . . . .	10
2.7.2. Метод максимальной правдоподобности . . . . .	10
2.8. Проверка гипотезы про распределение с помощью критерия Пирсона . . . . .	11
2.9. Нахождения доверительного интервала гипотетического закона распределения . . . . .	13
2.10. Выводы . . . . .	14

# 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 давайте построим дискретный вариационный ряд. Для этого строим табличку, в которой под каждой из вариантов содержится ее частота и частость.

варианта	0	1	2	3	4	5	6	7	8	9
частота $n_i$	1	10	20	18	20	15	9	2	1	4
частость $w_i$	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 представим в виде гистограммы, где каждый столбец представляет собой варианту, а высота каждого столбца - ее частота. Гистограмма имеет вид:

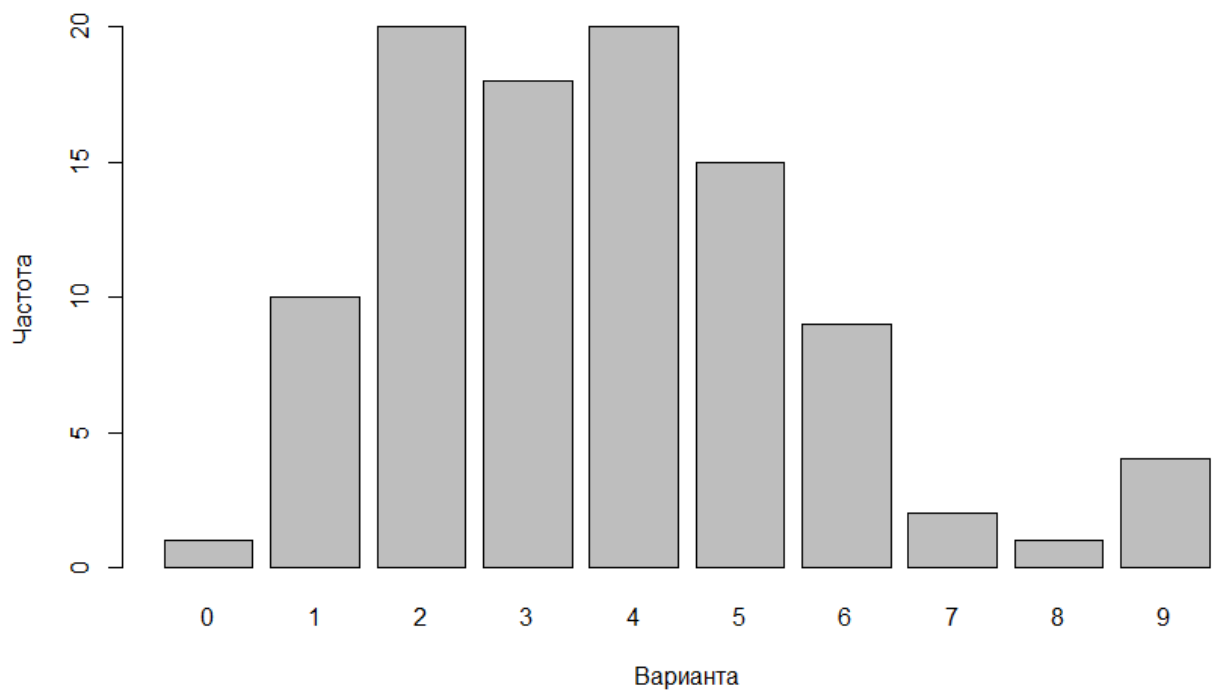


Рис. 1. Гистограмма выборки

## 2.3. Эмпирическая функция распределения

Построим эмпирическую функцию распределения:

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

Убрав промежуточные расчеты, получим

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

Ниже приводится график эмпирической функции распределения.

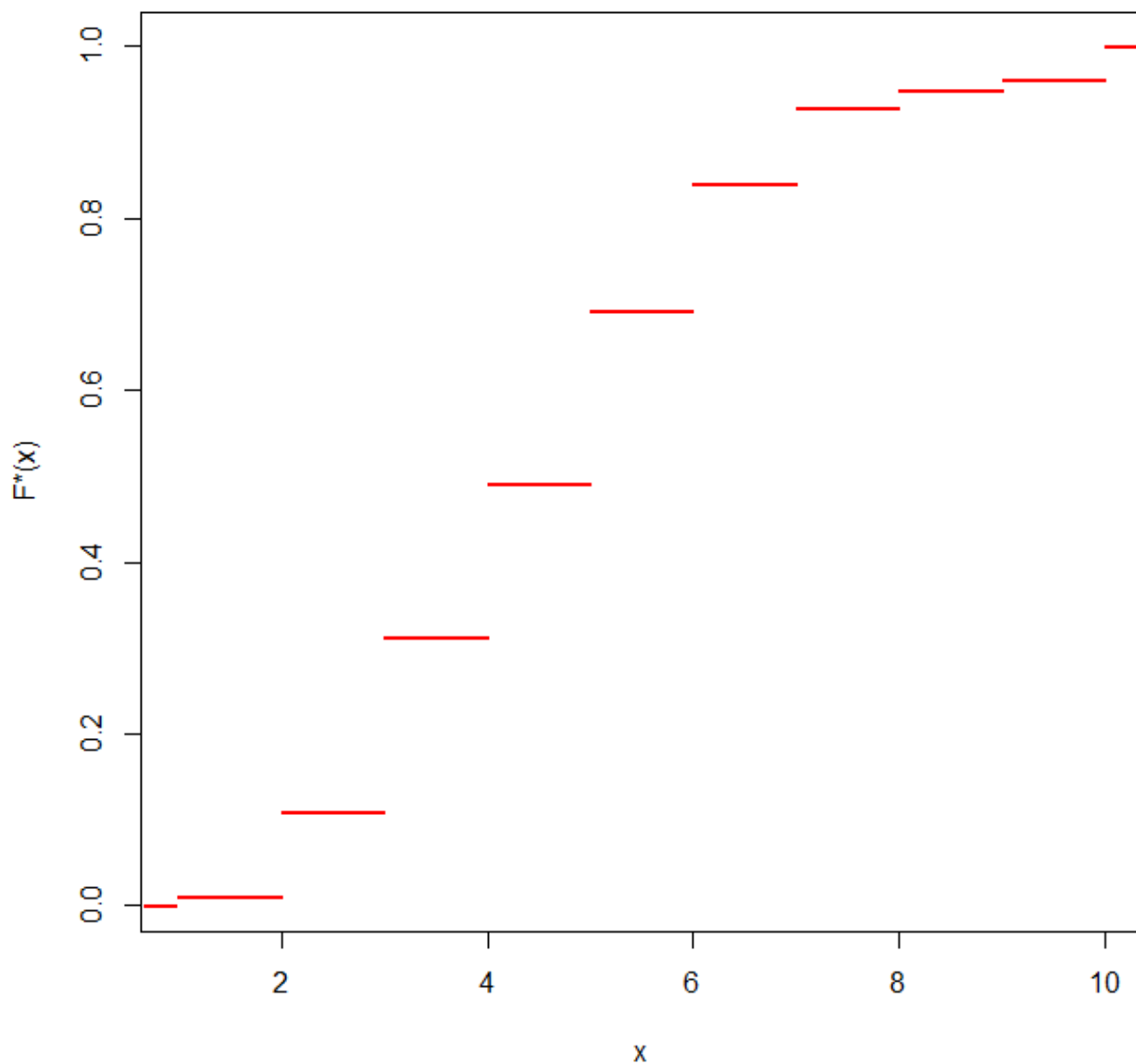


Рис. 2. График эмпирической функции распределения

## 2.4. Выборочные медиана, мода и асимметрия

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

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

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

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

Выборочная асимметрия считается по формуле

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

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

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

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

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

## 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

Мое предположение, что выборка порождена распределением Пуассона. Для этого есть несколько причин. Самое главное замечание, что в выборке содержатся 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 распределение. Также я поигрался с разными значениями параметров для распределения Пуассона и биномиального распределения. Ниже можно увидеть ряд графиков, которые подтверждают обе мои гипотезы.

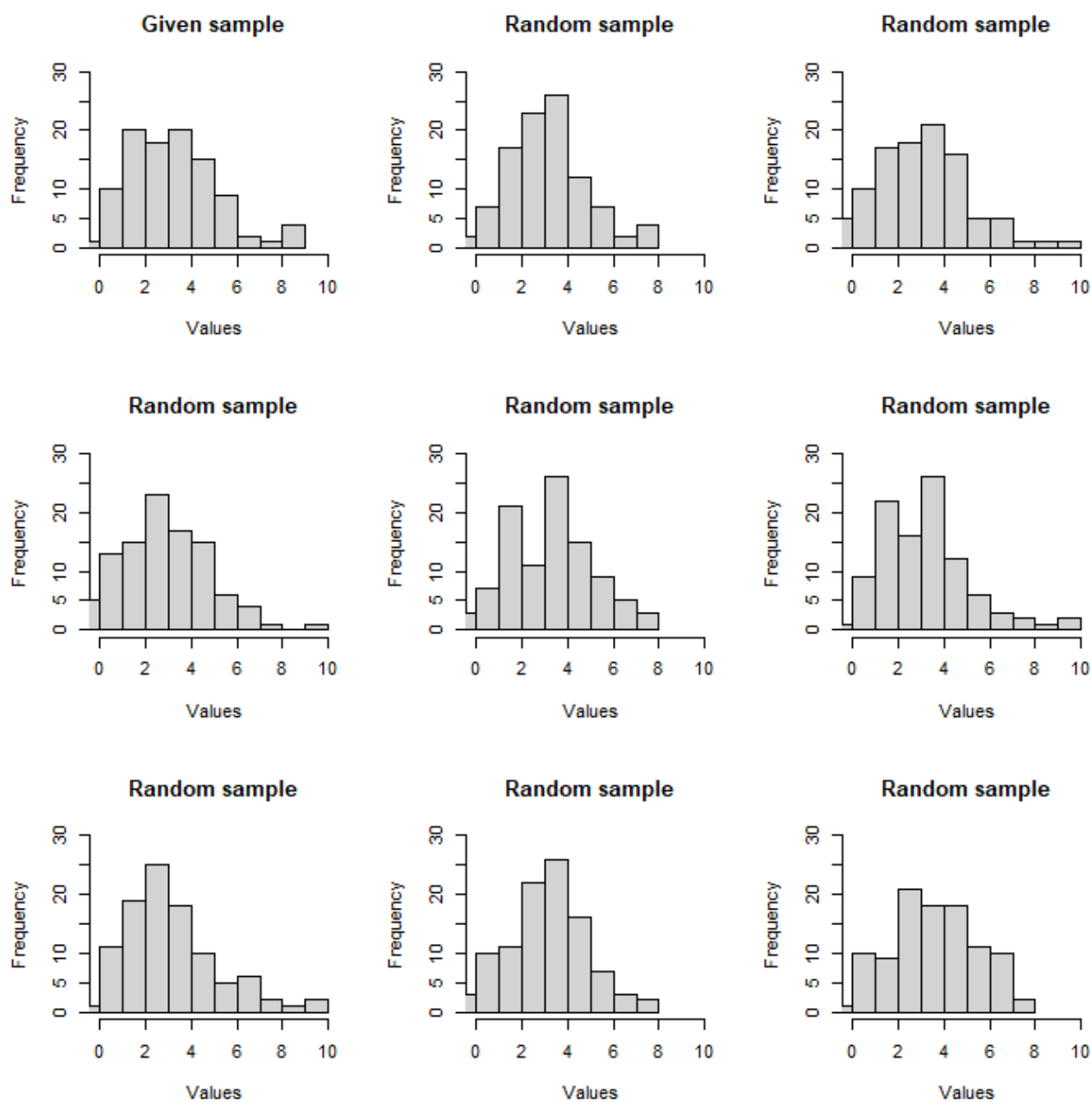


Рис. 3. Случайно сгенерированные выборки, порожденные законом Пуассона с параметром  $a = 3.71$



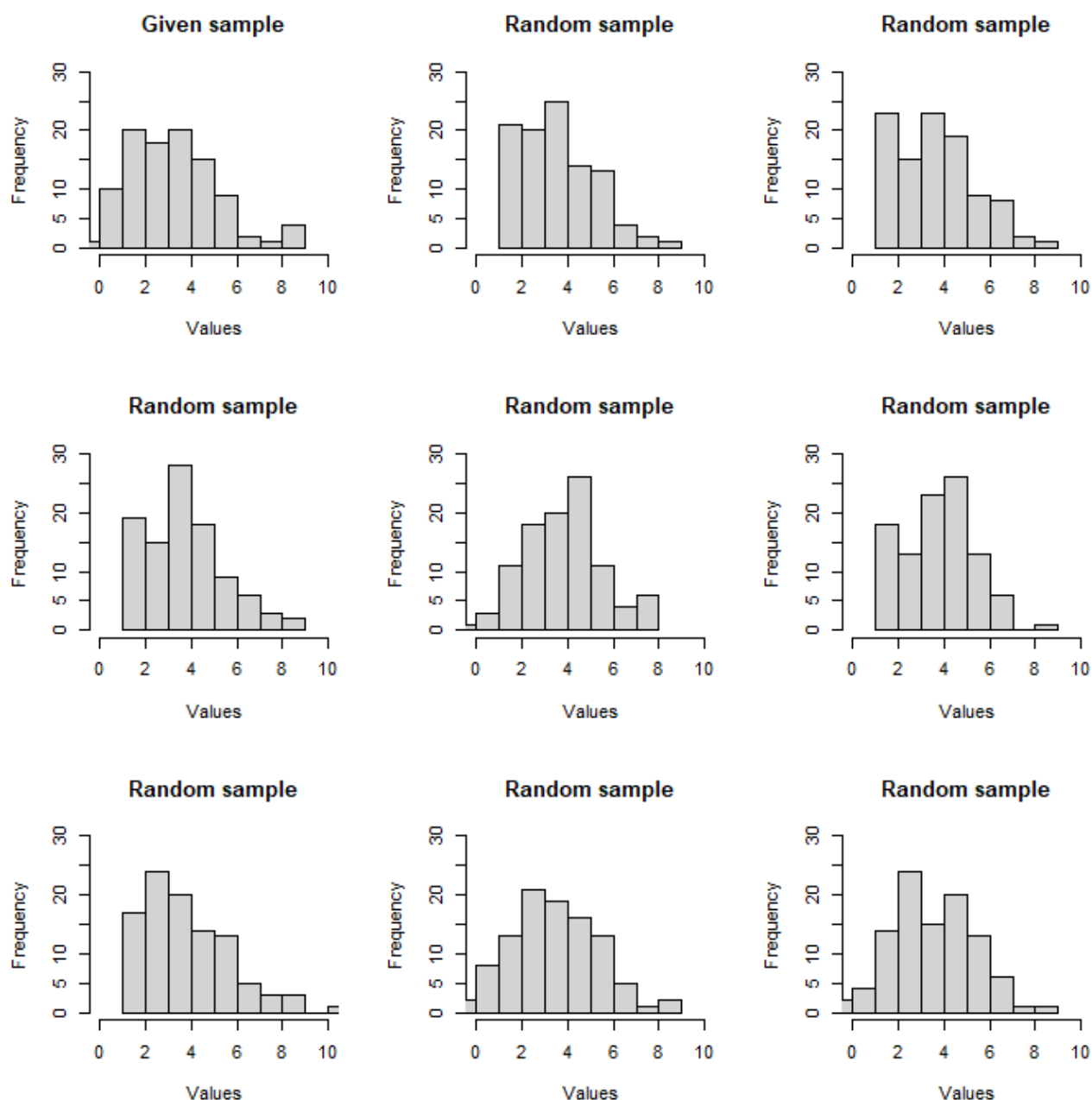


Рис. 4. Случайно сгенерированные выборки, порожденные биномиальным законом с параметрами  $size = 20, p = 0.2$

## 2.7. Нахождение точечной оценки параметра распределения Пуассона и проверка его свойств

Пусть наша выборка порождена распределением Пуассона, то есть

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

Известно, что математическое ожидание случайной величины, распределенной по закону 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 выборочное среднее есть состоятельной оценкой математического ожидания. Проверим эффективность:

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

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

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

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

### 2.7.1. Метод моментов

Суть метода моментов в том, что эмпирические моменты приравняются к теоретическим моментам. Если из этих равенств можно вытянуть числовое значение некой характеристики через неизвестные параметры, то круто. Эти значения можно брать за точечные оценки параметров. К делу, приравняем моменты первых порядков:

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

Получили, что  $a_{\text{ММ}}^* = \bar{x} = 3.71$ .

### 2.7.2. Метод максимальной правдоподобности

Суть метода максимальной правдоподобности в том, чтобы максимизировать значение функции правдоподобности. Прологарифмируем функцию правдоподобности распределения Пуассона:

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

Найдем критические точки:

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

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 стремиться к минус бесконечности в обоих случаях. А это значит, что в точке  $a_{cr}$  достигается максимум.

Получили, что  $a_{ММП}^* = \bar{x} = 3.71$ .

## 2.8. Проверка гипотезы про распределение с помощью критерия Пирсона

Пусть некая случайная величина распределена по закону Пуассона. Построим табличку, в которой будут храниться значения вероятностей, что эта случайная величина приняла 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	0.047	0.0218	0.0089

Сумма вероятностей в таблице есть

$$sum_{prob} = 0.9947.$$

Нам необходимо, чтобы сумма вероятностей была равна единице. Для этого вспомним, что случайная величина может принимать значения больше девяти. Добавим эту информацию в последний столбец. Тут мы немного пренебрежем округлением. Получим таблицу

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	$\geq 9$
вероятность	0.1433	0.0886	0.047	0.0218	0.0142

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

значение случ. вел.	0	1	2	3	4
$n \cdot p$	2.44	9.08	16.84	20.83	19.32
значение случ. вел.	5	6	7	8	9
$n \cdot p$	14.33	8.86	4.7	2.18	1.42

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

По алгоритму проверки гипотезы по критерию Пирсона, необходимо разбить множество возможных значений случайной величины на несколько непесекающихся классов. Поскольку случайная величина принимает десять возможных значений, то начальное число классов равно десяти. Наша цель - объединить некоторые классы так, чтобы в каждом из классов величина  $n \cdot p$  была не меньше десяти. Для этого объединим 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. $\xi_i$	$\{0, 1\}$	2	3	4	5	$\geq 6$
$n \cdot p_i$	11.52	16.84	20.83	19.32	14.33	17.16
частота $n_i$	11	20	18	20	15	16
разница $n_i - n \cdot p_i$	-0.52	3.16	-2.83	0.68	0.67	-1.16

Теперь необходимо подсчитать меру расхождения  $\eta_{val}$ . По определению

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

Подставляя числа из таблицы, получим

$$\eta_{val} = \frac{(-0.52)^2}{11.52} + \frac{(3.16)^2}{16.84} + \frac{(-2.83)^2}{20.83} + \frac{(0.68)^2}{19.32} + \frac{(0.67)^2}{14.33} + \frac{(-1.16)^2}{17.16} \approx 1.1346.$$

По теореме Пирсона статистика  $\eta$  стремится по распределению к распределению  $\chi^2_{r-s-1}$ , где  $r$  - число классов, а  $s$  - число неизвестных параметров. В нашем случае  $r = 6$  и  $s = 1$ , а уровень значимости  $\alpha = 0.05$  по условию. 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

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

Или

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

Или, нормируя случайную величину:

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

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  $\gamma$ . По условию  $\gamma = 0.95$ . Для упрощения задачи будем искать именно симметричный интервал. Тогда наша вероятность переписывается в виде

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

где  $t_\gamma$  находится из таблицы Лапласа. Поскольку

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

Запишем цепочку преобразований

$$\begin{aligned} |\eta| < 1.96 &\iff \frac{|\bar{x} - a| \cdot \sqrt{n}}{\sqrt{a}} < 1.96 \iff \\ \frac{|3.71 - a| \cdot 10}{\sqrt{a}} < 1.96 &\iff (3.71 - a)^2 \cdot 100 < 1.96^2 \cdot a \iff \\ a^2 - 7.42 \cdot a + 13.7641 < 0.038416 \cdot a &\iff 3.3512 < a < 4.10722. \end{aligned}$$

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

## 2.10. Выводы

Подведем итоги. Сначала было сделано предположение, что данная выборка порождена дискретным распределением. Тут сразу возникли сомнения на счет закона распределения. Поигравшись с графиками было обнаружено, что выборка похожа как и на  $Poiss(3.71)$ , так и на  $Bin(20, 0.2)$ . При этом значения параметров в биномиальном распределении брались на глаз. Сейчас можно немного объяснить их оправданность. Биномиальное распределение с такими параметрами имеет математическое ожидание  $E\xi = 4$  и дисперсию  $D\xi = 3.2$ . Как видим, эти величины похожи на аналогичные эмпирические. Больше уверенности приходит, если высчитать медиану  $Me\xi = 4$  и моду  $Mo\xi = 4$ . А вот с асимметрией  $As\xi \cong 0.33$  есть расхождения, ведь  $As^*\xi \cong 0.709$ . Но это неудивительно: во-первых, расхождение не так уж и велико; во-вторых, асимметрия показывает, насколько график асимметричен. Понятно, что в какой-то конкретной выборке может случиться перевес в одну из сторон, пусть и небольшой. Это могло и случиться с нашей выборкой, поэтому из-за такой нестыковки в асимметрии нельзя сразу же отбрасывать закон. По хорошему, нужно проводить вычисления, аналогичные тому, что уже приведены в работе. Поскольку это немного затратное занятие, то в работе нет ни единой проверки на возможность биномиального распределения. Но все же интерес берет верх, поэтому я сделал поверхностные быстрые расчеты.

По их результатам выборка могла быть порождена биномиальным распределением. Точнее, она не противоречит критерию Пирсона с уровнем значимости  $\alpha = 0.95$ .

Что касается всех проверок на распределение Пуассона, то здесь проблем не возникло. Не было никаких сомнений, что что-то расходится или не складывается. Единственное мое удивление было, когда при проверке на удовлетворение критерия Пирсона, у меня значение меры расхождения вышло намного меньшим, чем критическое пороговое значение. У меня была идея увеличить количество интервалов, пусть при этом не все рекомендательные условия выполнялись бы. Но я все же решил этого не делать. Действительно, по моим расчетам на глаз, мера расхождения не превышала бы критическое пороговое значение.