Курсовая работа

Идентификация законов распределения случайных величин

Пусть в (статистическом) эксперименте доступна наблюдению случайная величина X, распределение которой P неизвестно полностью или частично. Тогда любое утверждение, касающееся P называется ***статистической гипотезой***. Гипотезы различают по виду предположений, содержащихся в них:

– Статистическая гипотеза, однозначно определяющая распределение P, то есть $H:\{P=P_0\}$, где P_0 какой-то конкретный закон, называется ***простой*** (известен закон распределения вплоть до параметров).

– Статистическая гипотеза, утверждающая принадлежность распределения P к некоторому семейству распределений, то есть вида $H:\{P\in\rho\}$, где ρ- семейство распределений, называется **сложной** (неизвестные параметры предполагаемого закона находятся по статистической выборке).

Статистической гипотезой называется любое предположение о виде или параметрах неизвестного закона распределения.

Проверка статистической гипотезы

H_0 – ошибка первого рода, имеет место, когда отвергается верная гипотеза;

H_1 (альтернативная гипотеза) – ошибка второго рода – принимается неверная гипотеза.

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (ошибка второго рода)
	H_1	H_0 неверно отвергнута (ошибка первого рода)	H_0 верно отвергнута

Как видно из вышеприведённого определения, ошибки первого и второго рода являются взаимно-симметричными, то есть если поменять местами гипотезы H_0 и H_1, то ошибки первого рода превратятся в ошибки второго рода и наоборот. Тем не менее, в большинстве практических ситуаций путаницы не про-

исходит, поскольку принято считать, что нулевая гипотеза H_0 соответствует состоянию «по умолчанию» (естественному, наиболее ожидаемому положению вещей) — например, что обследуемый человек здоров, или что проходящий через рамку металлодетектора пассажир не имеет запрещённых металлических предметов. Соответственно, альтернативная гипотеза H_1 обозначает противоположную ситуацию, которая обычно трактуется как менее вероятная, неординарная, требующая какой-либо реакции.

Вероятность ошибки первого рода при проверке статистических гипотез называют уровнем значимости и обычно обозначают греческой буквой **α** (отсюда название **α-errors**).

Вероятность ошибки второго рода не имеет какого-то особого общепринятого названия, на письме обозначается греческой буквой **β** (отсюда **β-errors**). Однако с этой величиной тесно связана другая, имеющая большое статистическое значение — **мощность критерия**. Она вычисляется по формуле (1−β). Таким образом, чем выше мощность, тем меньше вероятность совершить ошибку второго рода.

Наиболее часто применяемые *уровни значимости*: 0,2; 0,1; 0,05; 0,01.

Критерий согласия (статистический критерий) — строгое математическое правило, по которому принимается или отвергается та или иная статистическая гипотеза с известным уровнем значимости. Построение критерия представляет собой выбор подходящей функции от результатов наблюдений (ряда эмпирически полученных значений признака), которая служит для выявления меры расхождения между эмпирическими значениями и гипотетическими.

Наиболее известные критерии согласия:

- χ^2 (хи-квадрат) Пирсона;
- типа Колмогорова-Смирнова;
- Андерсона-Дарлинга;
- омега-квадрат Мизеса

и т.д.

Критерий χ^2 (хи-квадрат) Пирсона.

Наблюдаемое значение критерия находится по формуле:

$$\chi^2_{набл.} = \sum_{i=1}^{S} \frac{(M_i - M_i^{'})^2}{M_i^{'}},$$

где M_i и $M_i^{'}$ – соответственно эмпирические и теоретические частоты распределения.

S – число интервалов статистического ряда, построенного по данным выборки.

Число степеней свободы распределения $k = S - r - 1$,

где S – число интервалов, r – число параметров предполагаемого теоретического распределения.

Для проверки гипотезы строится критическая область.

Размер статистической выборки (выборки значений случайной величины) для проверки гипотезы выбирается произвольно и равен N.

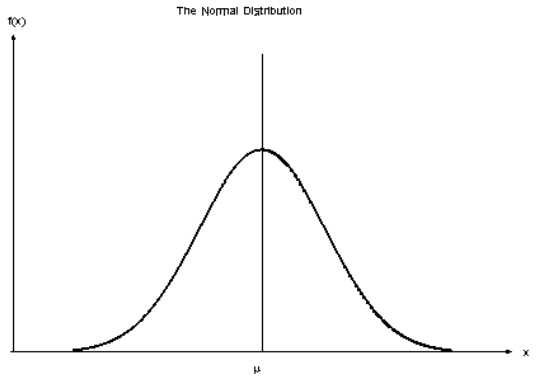
Число интервалов выборки определяется по формуле Стерджесса:

$$S = 1 + 3{,}322 \lg N,$$

где N – число единиц совокупности (значений случайной величины).

Наиболее распространенные законы распределения:

– нормальный закон (теорема Чебышева или закон больших чисел);



– закон Вейбулла (отказы оборудования в теории надежности);

– закон Бернулли (моделирует случайный эксперимент произвольной природы, когда заранее известны вероятности успеха или неудачи, принимает всего два значения – 0 и 1);

– экспоненциальный (показательный) закон;

– закон Пуассона (моделирует случайную величину, представляющую собой число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга);

В общем случае обработка статистического распределения состоит из следующей последовательности этапов:

1) Построение статистического ряда.

2) Выдвижение гипотезы о предполагаемом теоретическом распределении, соответствующем эмпирическим данным.

3) Определение параметров теоретического распределения.

4) Проверка согласия предполагаемого теоретического и эмпирического распределений по определенному критерию (критериям).

Для определения закона распределения некоторой случайной величины рассматриваются статистические данные, собранные по достаточно большому количеству независимых наблюдений.

Необходимо преобразовать дискретную выборку случайных чисел в интервальный ряд, рассчитать частоты попадания данной случайной величины в полученные интервалы и определить числовые характеристики эмпирического распределения.

Шаг (величина интервала) определяется следующим образом:

$$i = \frac{x_{\max} - x_{\min}}{S},$$

где x_{max} – максимальное значение случайной величины;

x_{min} – минимальное значение случайной величины;

S – число интервалов.

Для построения теоретической кривой распределения случайной величины необходимо найти некоторые числовые характеристики теоретического распределения. По заданному эмпирическому распределению вычисляется выборочная средняя $\overline{X}$, дисперсия D и среднеквадратическое отклонение *СКО*.

Выборочная средняя находится по формуле:

$$\overline{X} = \frac{\sum_i \overline{X_i} \cdot M_i}{\sum_i M_i},$$

где M_i - частота попадания случайной величины в интервал.

Дисперсия вычисляется следующим образом:

$$D(X) = \frac{\sum_i (X_i - \overline{X})^2 \cdot M_i}{\sum_i M_i}.$$

Среднеквадратическое отклонение определяется по формуле:

$$\sigma = \sqrt{D(X)}.$$

Частость i-того интервала вычисляется по формуле:

Рассмотрим примеры определения некоторых законов распределения случайных величин.

Закон Пуассона

Определим закон распределения некоторой случайной величины. Выдвигается нулевая гипотеза H_0 о том, что расхождение эмпирических частот распределения случайной величины и теоретических частот распределения Пуассона незначимо (генеральная совокупность случайных чисел распределена по закону Пуассона).

Объем выборки (N) – 100.

Число интервалов (I) – 8.

Максимальное значение случайной величины (x_{max}) – 8.

Минимальное значение случайной величины ($x_{\min}$) – 1.

Шаг (величина интервала) (i) – 1.

Числовые характеристики эмпирического распределения сведены в таблицу 1.

Таблица 1 – Эмпирическое распределение случайной величины и его числовые характеристики

№	Нижняя граница X_i	Верхняя граница X_{i+1}	Частота M_i	Частость W_i	Центр интервала $\overline{X_i}$	$M_i \cdot \overline{X_i}$	Отклонение от среднего $\overline{X_i} - \overline{X}$	Квадрат отклонения $(X_i - \overline{X})^2$	$M_i(X_i - \overline{X})^2$
1	1	1	2,00	0,02	1,00	2,00	-3,33	11,09	22,18
2	2	2	8,00	0,08	2,00	16,00	-2,33	5,43	43,43
3	3	3	20,00	0,20	3,00	60,00	-1,33	1,77	35,38
4	4	4	30,00	0,30	4,00	120,00	-0,33	0,11	3,27
5	5	5	18,00	0,18	5,00	90,00	0,67	0,45	8,08
6	6	6	12,00	0,12	6,00	72,00	1,67	2,79	33,47
7	7	7	7,00	0,07	7,00	49,00	2,67	7,13	49,90
8	8	8	3,00	0,03	8,00	24,00	3,67	13,47	40,41
Итого			100	1,00		433			236,11

Числовые характеристики теоретического распределения:

- выборочная средняя $\overline{X} = 4{,}33$;

- дисперсия $D(X) = 2{,}3611$;

- среднеквадратическое отклонение $\sigma = 1{,}544$.

Для более наглядного представления построим полигон эмпирического распределения рассматриваемой случайной величины (рисунок 1).

Рисунок 1 – Полигон эмпирического распределения случайной величины

Вид полученной гистограммы, численная близость среднего и дисперсии, а также характер исследуемого потока (события происходят с постоянной средней интенсивностью) позволяют предположить, что исследуемая величина подчиняется закону Пуассона.

Для того, чтобы при уровне значимости α проверить гипотезу о распределении генеральной совокупности по закону Пуассона, необходимо:

1) принять в качестве оценки параметра λ закона Пуассона величину, равную выборочной средней:

$$\lambda = \overline{X} = 4{,}33;$$

2) найти вероятности попадания случайной величины в частичные интервалы ($X_i; X_{i+1}$) по формуле:

$$P_i = (X_i < X < X_{i+1}) = \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} - \frac{\lambda^{X_{i+1}}}{X_{i+1}!} e^{-\lambda};$$

3) найти теоретические частоты по формуле:

$$M_i^{'} = P_i \cdot \sum M_i;$$

4) сравнить эмпирические и теоретические частоты с помощью критерия Пирсона, приняв число степеней свободы $K = S - 1$ и сделать вывод о достоверности гипотезы:

$$\chi^2_{набл.} = \sum_{i=1}^{S} \frac{(M_i - M_i^{'})^2}{M_i^{'}},$$

где M_i - практическое число попаданий случайной величины в i-ый интервал;

$M_i^{'}$ – теоретическое число значений в i-ом интервале;

S – число интервалов статистического ряда, построенного по данным выборки.

Результаты расчета наблюдаемого значения критерия Пирсона поместим в таблицу 2.

Таблица 2 – Вычисление наблюдаемого значения χ^2

№	M_i	P_i	$M_i^{'}$	$M_i - M_i^{'}$	$(M_i - M_i^{'})^2$	$\frac{(M_i - M_i^{'})^2}{M_i^{'}}$
1	2,00	0,05702	5,7015	-3,7015	13,70	2,40311
2	8,00	0,12344	12,3439	-4,3439	18,87	1,52862
3	20,00	0,17816	17,8163	2,1837	4,77	0,26765
4	30,00	0,19286	19,2861	10,7139	114,79	5,95178
5	18,00	0,16702	16,7018	1,2982	1,69	0,10091
6	12,00	0,12053	12,0531	-0,0531	0,00	0,00023
7	7,00	0,07456	7,4557	-0,4557	0,21	0,02786
8	3,00	0,04035	4,0354	-1,0354	1,07	0,26567
Итого						10,54582

Используя таблицу критических точек распределения χ^2, при числе степеней свободы, равном 7 (число разрядов S равно 8, а число связей равно 1, так как распределение Пуассона оценивается одним параметром - λ), и уровне значимости $\alpha = 0{,}05$ (то есть вероятность расхождения теоретического распределения с эмпирическим распределением меньше 0,05, и вероятность соответствия его закону Пуассона больше 0,95) критическое значение равно $\chi^2_{\text{крит}} = 10{,}8$.

Таким образом, выдвинутая гипотеза принимается, так как: $\chi^2_{набл} = 10{,}5 < \chi^2_{\text{крит}} = 10{,}8$.

В результате обработки статистического материала было получено, что рассматриваемую случайную величину можно аппроксимировать пуассоновской случайной величиной с параметром $\lambda = 4{,}33$.

Нормальный закон

Определим закон распределения некоторой случайной величины. Выдвигается нулевая гипотеза H_0 о том, что расхождение эмпирических частот распределения случайной величины и теоретических частот нормального распределения незначимо (генеральная совокупность случайных чисел распределена по нормальному закону).

Объем выборки (N) – 108.

Число интервалов (I) – 8.

Максимальное значение случайной величины ($x_{\max}$) – 2687,6.

Минимальное значение случайной величины ($x_{\min}$) – 702,8.

Шаг (величина интервала) (i) – 248,1.

Числовые характеристики эмпирического распределения сведены в таблицу 3.

Таблица 3 – Эмпирическое распределение случайной величины и его числовые характеристики

№	Нижняя граница X_i	Верхняя граница X_{i+1}	Частота M_i	Частость W_i	Центр интервала $\overline{X_i}$	$M_i \cdot \overline{X_i}$	Отклонение от среднего $\overline{X_i} - \overline{X_в}$	Квадрат отклонения $(X_i - \overline{X_i})^2$	$M_i(X_i - \overline{X_i})^2$
1	702,8	950,9	6	0,06	826,9	4961,1	-902,8	815062,9	4890377,3
2	950,9	1199,0	14	0,13	1075,0	15049,3	-654,7	428643,0	6001002,0
3	1199,0	1447,1	17	0,16	1323,1	22491,8	-406,6	165330,3	2810615,7
4	1447,1	1695,2	27	0,25	1571,2	42421,1	-158,5	25124,9	678372,1
5	1695,2	1943,3	21	0,19	1819,3	38204,3	89,59	8026,66	168561,0
6	1943,3	2191,4	15	0,14	2067,4	31010,3	337,7	114035,7	1710534,9
7	2191,4	2439,5	5	0,04	2315,5	11577,3	585,8	343151,9	1715759,4
8	2439,5	2687,6	3	0,03	2563,6	7690,7	833,9	695375,3	2086125,9
Итого			108	1		173405,9			20061348,3

Выборочная средняя: $\overline{X_в} = 1605{,}6$.

Дисперсия: $D(X) = 20061{,}35$.

Среднеквадратическое отклонение: $\sigma = 447{,}899$.

Построим гистограмму эмпирического распределения (рисунок 2).

Рисунок 2 – Гистограмма эмпирического распределения

Вид полученной гистограммы (рисунок 2), численная близость среднего и математического ожидания, а также характер исследуемого потока (события формируются под действием большого числа независимых факторов) позволяют предположить, что исследуемая величина подчиняется нормальному закону.

Результаты вычисления теоретической вероятности сведены в таблицу 4.

Таблица 4 – Вычисление теоретических вероятностей попадания в заданный интервал нормально распределенной случайной величины

Номер	X_i	X_{i+1}	$\frac{X_{i+1}-\overline{X_в}}{\sigma}$	$\frac{X_i-\overline{X_в}}{\sigma}$	$\Phi(\frac{X_{i+1}-\overline{X_в}}{\sigma})$	$\Phi(\frac{X_i-\overline{X_в}}{\sigma})$	P_i
1	702,8	950,9	-1,4617	-2,0157	-0,4279	-0,4783	0,0504
2	950,9	1199,0	-0,9078	-1,4617	-0,3186	-0,4279	0,1093
3	1199,0	1447,1	-0,3539	-0,9078	-0,1368	-0,3186	0,1818
4	1447,1	1695,2	0,2000	-0,3539	0,0793	-0,1368	0,2161
5	1695,2	1943,3	0,7539	0,2000	0,2734	0,0793	0,1941
6	1943,3	2191,4	1,3079	0,7539	0,4049	0,2734	0,1315
7	2191,4	2439,5	1,8618	1,3079	0,4686	0,4049	0,0637
8	2439,5	2687,6	2,4157	1,8618	0,4922	0,4686	0,0236
Итого							0,9705

Таблица интегральной функции Лапласа приведена в приложении А.

Результаты расчета наблюдаемого значения критерия Пирсона представлены в таблице 5.

Таблица 5 – Вычисление наблюдаемого значения χ^2

Номер	M_i	P_i	$M_i^{'}$	$M_i - M_i^{'}$	$(M_i - M_i^{'})^2$	$\frac{(M_i - M_i^{'})^2}{M_i^{'}}$
1	6	0,0504	5,443	0,5568	0,31	0,0569
2	14	0,1093	11,8	2,1956	4,82	0,4084
3	17	0,1818	19,63	-2,6344	6,94	0,3535
4	27	0,2161	23,34	3,6612	13,40	0,5743
5	21	0,1941	20,96	0,0372	0,001	0,000066
6	15	0,1315	14,2	0,798	0,64	0,0448
7	5	0,0637	8,73	-3,7269	13,89	1,5916
8	3	0,0236	2,54	0,4512	0,20	0,0799
Итого	108	0,9705	97,05			3,1095

Используя таблицу критических точек распределения χ^2, найдём при числе степеней свободы $k = 8 - 3 = 5$ (число разрядов равно 8, а число связей S равно 3, так как нормальное распределение оценивается двумя параметрами – математическим ожиданием и среднеквадратическим отклонением) и уровне значимости $\alpha = 0{,}1$ (то есть вероятность расхождения теоретического

распределения с нормальным распределением меньше 0,1, и вероятность соответствия его нормальному закону больше 0,9) критическую точку: $\chi^2_{\text{крит}} = 9{,}2$. Выдвинутая гипотеза принимается, так как: $\chi^2_{набл} = 3{,}11 < \chi^2_{\text{крит}} = 9{,}2$.

Таким образом, в целом, несмотря на ограниченный объем выборки эмпирических данных, можно предположить, что распределения рассматриваемой случайной величины близко к нормальному закону (хотя эта гипотеза проверялась при небольшой вероятности).

В результате обработки статистического материала было получено, что случайную величину можно аппроксимировать нормально распределенной случайной величиной с параметрами: математическим ожиданием $\overline{X_в} = 1605{,}6$; среднеквадратическим отклонением $\sigma = 447{,}899$.

Показательный (экспоненциальный) закон

Определим закон распределения некоторой случайной величины. Выдвигается нулевая гипотеза H_0 о том, что расхождение эмпирических частот распределения случайной величины и теоретических частот показательного распределения незначимо (генеральная совокупность случайных чисел распределена по показательному (экспоненциальному) закону).

Объем выборки (N) – 100.

Число интервалов (I) – 8.

Максимальное значение случайной величины ($x_{\max}$) – 12.

Минимальное значение случайной величины ($x_{\min}$) – 181,92.

Шаг (величина интервала) (i) – 21,24.

Числовые характеристики эмпирического распределения сведены в таблицу 6.

Таблица 6 – Эмпирическое распределение случайной величины и его числовые характеристики

№	Нижняя граница X_i	Верхняя граница X_{i+1}	Частота M_i	Частость W_i	Центр интервала $\overline{X_i}$	$M_i \cdot \overline{X_i}$	Отклонение от среднего $\overline{X_i} - \overline{X_в}$	Квадрат отклонения $(X_i - \overline{X_i})^2$	$W_i(X_i - \overline{X_i})^2$
1	12,00	33,24	29	0,29	22,62	655,98	-53,9488	2910,47	84403,72
2	33,24	54,48	21	0,21	43,86	921,06	-32,7088	1069,87	22467,18
3	54,48	75,72	15	0,15	65,1	976,5	-11,4688	131,53	1973,00
4	75,72	96,96	12	0,12	86,34	1036,08	9,7712	95,48	1145,72
5	96,96	118,20	9	0,09	107,58	968,22	31,0112	961,69	8655,25
6	118,20	139,44	6	0,06	128,82	772,92	52,2512	2730,19	16381,13
7	139,44	160,68	5	0,05	150,05	750,25	73,4912	5400,96	27004,78
8	160,68	181,92	3	0,03	171,29	513,87	94,7112	8970,21	26910,63
Итого			100	1,0		6594,88			188941,41

Выборочная средняя: $\overline{X_в} = 65{,}9488$.

Дисперсия: $D(X) = 1889{,}41407$.

Среднеквадратическое отклонение: $\sigma = 43{,}4674$.

Параметр: $\lambda = 0{,}01516$.

Построим гистограмму эмпирического распределения (рисунок 3).

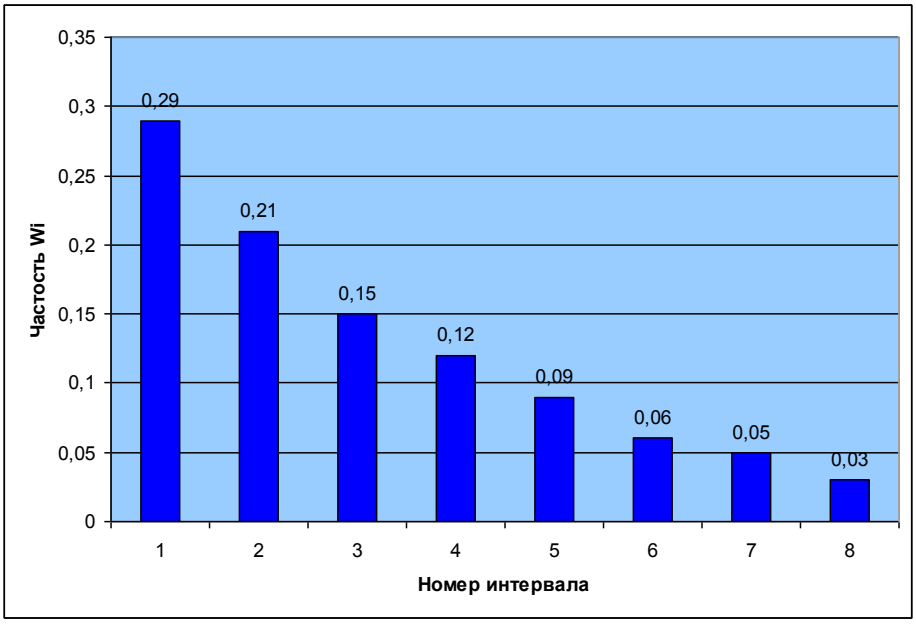


Рисунок 3 – Гистограмма эмпирического распределения

Вид полученной гистограммы (рисунок 3), численная близость среднего и математического ожидания, позволяют предположить, что исследуемая величина подчиняется показательному (экспоненциальному) закону.

Результаты вычисления теоретической вероятности сведены в таблицу 7.

Таблица 7 - Вычисление теоретических вероятностей

№	Интервал		$-\lambda x_i$	$-\lambda x_{i+1}$	$e^{-\lambda x_i}$	$e^{-\lambda x_{i+1}}$	P_i
	x_i	x_{i+1}					
1	12,00	33,24	-0,1818	-0,5037	0,8338	0,6043	0,2294
2	33,24	54,48	-0,5037	-0,8255	0,6043	0,4380	0,1663
3	54,48	75,72	-0,8255	-1,1474	0,4380	0,3175	0,1205
4	75,72	96,96	-1,1474	-1,4692	0,3175	0,2301	0,0874
5	96,96	118,20	-1,4692	-1,7911	0,2301	0,1668	0,0633
6	118,20	139,44	-1,7911	-2,1129	0,1668	0,1209	0,0459
7	139,44	160,68	-2,1129	-2,4347	0,1209	0,0876	0,0333
8	160,68	181,92	-2,4347	-2,8020	0,0876	0,0607	0,0269
Σ							0,773

Таблица 8 - Вычисление наблюдаемого значения критерия $\chi^2_{набл}$

N	M_i	P_i	$n \cdot P_i$	$M_i - n \cdot P_i$	$(M_i - n \cdot P_i)^2$	$\frac{(M_i - n \cdot P_i)^2}{n \cdot P_i}$
1	29	0,2294	22,942	6,0579	36,6988	1,5996
2	21	0,1663	16,6292	4,3708	19,1039	1,1488
3	15	0,1205	12,0534	2,9466	8,6823	0,7203
4	12	0,0874	8,7367	3,2633	10,6488	1,2189
5	9	0,0633	6,3327	2,6673	7,1145	1,1235
6	6	0,0459	4,5902	1,4098	1,9876	0,4330
7	5	0,0333	3,3241	1,6729	2,7986	0,8411
8	3	0,0269	2,6939	0,3061	0,0937	0,0348
	100	0,773	77,3053			7,12

Используя таблицу критических точек распределения χ^2, найдём при числе степеней свободы $k = 8 - 2 = 6$ (число разрядов равно 8, а число связей S равно 2, так как показательное распределение оценивается одним параметром – λ) и уровне значимости $\alpha = 0{,}05$ (то есть вероятность расхождения теоретического распределения с нормальным распределением меньше 0,05, и вероятность соответствия его показательному закону больше 0,95) критическую точку: $\chi^2_{\text{крит}} = 12{,}6$. Выдвинутая гипотеза принимается, так как: $\chi^2_{набл} = 7{,}12 < \chi^2_{\text{крит}} = 12{,}6$.

Таким образом, в целом, несмотря на ограниченный объем выборки эмпирических данных, можно предположить, что распределения рассматриваемой случайной величины близко к показательному закону.

В результате обработки статистического материала было получено, что случайную величину можно аппроксимировать показательно распределенной случайной величиной с параметром $\lambda = 0{,}01516$.

Задание на курсовую работу:

1) Собрать статистические данные по двум случайным величинам, выделенным в ходе анализа выбранного бизнес-процесса.

2) Провести идентификацию законов распределения случайных величин.