15. Основы математической статистики

Математическая статистика – это раздел математики, в котором изучаются математические методы планирования экспериментов, систематизации, обработки и использования статистических данных для научных и практических целей. В математической статистике предполагается, что результаты опытных данных и наблюдений являются реализацией случайных величин или процессов, имеющих те или иные законы распределения. Методы математической статистики обосновывают способы группировки и анализа статистических сведений о качественных и количественных признаках объектов различной природы. Проведение обследования каждого объекта большой совокупности относительно интересующего признака или физически невозможно или экономически нецелесообразно. Для установления статистических закономерностей случайно отбирают из всей совокупности ограниченное число объектов и подвергают их изучению.

15.1. Числовые характеристики статистических распределений

Совокупность всех подлежащих изучению объектов называется *генеральной совокупностью*. *Выборочной совокупностью (выборкой)* называется совокупность объектов, отобранных случайным образом из генеральной совокупности.

Число объектов совокупности (выборочной или генеральной) называется ее *объемом n*.

Простым случайным называют такой отбор, при котором заранее пронумерованные объекты извлекают из генеральной совокупности по номеру, который определяют из таблицы случайных чисел. Если оказалась, что случайное число из таблицы превышает число объектов генеральной совокупности *N,* то такое случайное число пропускают.

Если при формировании выборки пропускают ранее встречавшиеся числа из таблицы случайных чисел, то такая выборка называется *бесповторной*, если не пропускают, то получается повторный отбор, а выборка – повторной.

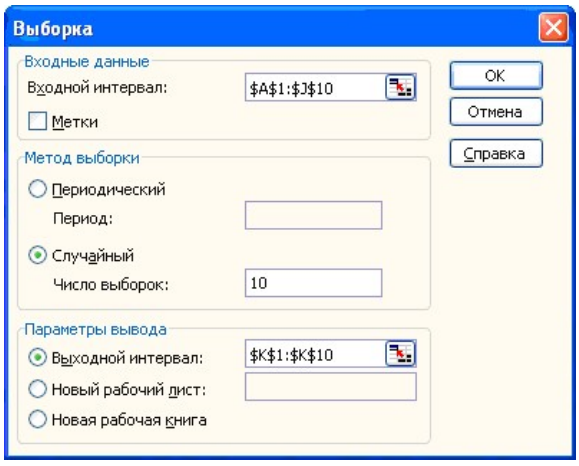
Подпрограмма «Выборка» из пакета «Анализ данных» электронного процессора Excel в автоматическом режиме осуществляет случайный отбор чисел из генеральной совокупности.

Пример. Необходимо сформировать выборку объемом 10 чисел из генеральной совокупности объемом 100 чисел:

Генеральная совокупность										
№	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
1	39	64	67	99	37	63	49	55	100	44
2	41	67	80	77	54	64	48	65	60	74
3	61	56	59	71	28	50	53	50	56	71
4	65	13	33	59	69	53	58	63	113	28
5	40	73	55	16	63	34	33	38	37	23
6	43	73	56	61	24	74	89	66	62	80
7	42	57	27	51	61	53	59	73	58	68
8	61	69	32	68	41	73	57	69	80	64
9	38	50	40	70	66	59	71	41	85	84
10	58	84	16	50	49	68	87	98	78	72

Решение

Сначала разместим все числа генеральной совокупности в ячейках $A$1:$J$10. Затем вызваем диалоговое окно подпрограммы «Выборка» (*Сервис→ Надстройки → Анализ данных → Выборка*), введем адреса ячеек генеральной совокупности, информацию об объеме формируемой выборке и адреса ячеек размещения случайно отобранных чисел $K$1:$K$10 (рисунок справа).



В итоге исходная таблица дополняется результатом – полученной выборкой в указанном диапазоне (сами выбранные значения выделены в таблице).

Генеральная совокупность											Выборка
№	A	B	C	D	E	F	G	H	I	J	K
1	39	64	67	99	37	63	49	55	100	44	64
2	41	67	80	77	54	64	48	65	60	74	57
3	61	56	59	71	28	50	53	50	56	71	78
4	65	13	33	59	69	53	58	63	113	28	57
5	40	73	55	16	63	34	33	38	37	23	77
6	43	73	56	61	24	74	89	66	62	80	66
7	42	57	27	51	61	53	59	73	58	68	64
8	61	69	32	68	41	73	57	69	80	64	69
9	38	50	40	70	66	59	71	41	85	84	72
10	58	84	16	50	49	68	87	98	78	72	54

Серийным называют отбор, при котором объекты извлекают из генеральной совокупности не по одному, а «сериями», которые полностью подвергаются исследованию. Серийный отбор применяется тогда, когда изменения изучаемого признака в различных сериях незначительны.

Часто на практике используют *комбинированный* отбор, при котором сочетаются различные способы. Иногда генеральную совокупность разбивают на серии одинакового объема, а затем простым случайным отбором выбирают несколько серий, из каждой серии простым случайным отбором извлекают отдельные объекты.

Операция расположения значений случайной величины по возрастанию (не убыванию) называется *ранжированием статистических данных.* Полученная таким образом последовательность значений случайной величины X называется *вариационным рядом*. Сами значения x_i случайной величины называются *вариантами.*

Числа n_i, показывающие сколько раз встречаются варианты x_i в вариационном ряде, называются ***частотами*** $\left(\sum_{i=1}^{k} n_i = n\right)$, а их отношение к объему выборки – *относительными частотами* $w_i = \frac{n_i}{n}$ $\left(\sum_{i=1}^{k} w_i = 1\right)$, где k – количество вариант.

Перечень вариантов и соответствующих им частот или относительных частот называется *вариационным рядом* или *статистическим распределением.*

Интервальным статистическим распределением выборки называют перечень интервалов $(x_i; x_{i+1})$ и соответствующих им частот n_i или относительных частот $w_i = n_i / n$ (в качестве частоты n_i, соответствующей интервалу, принимают количество случайных чисел, попавших в этот интервал).

Если в качестве представителя каждого интервала выбрать значение его середины $x_i^* = (x_i + x_{i+1})/2$ и составить перечень соответствующих частот n_i или относительных частот w_i, то в этом случае можно получить вариационный ряд интервального статистического распределения.

Распределение чисел из выборки по интервалам $(x_i; x_{i+1})$ становится возможным, если определена его длина h, вычисляемая по формуле $h = \frac{R}{k}$, $R = X_{\max} - X_{\min}$, где R – размах выборки, $X_{\max}$ и $X_{\min}$ – максимальное и минимальное значение объекта выборки, k – количество интервалов.

Желательно перед вычислением длины интервала h максимальное $X_{\max}$ и минимальное $X_{\min}$ значения выборки округлить до целых значений ($X_{\max}$ округляется до целого в сторону увеличения, $X_{\min}$ – в сторону уменьшения).

Количество интервалов k (целое число) целесообразно выбрать не менее 7, но и не более 15 или определить по формуле Старджесса $k = 1 + 3{,}322 \cdot \lg n$, где n – объем выборки. Если k, вычисляемое по формуле Старджесса, нецелое число, то в качестве числа интервалов надо взять ближайшее к k целое число, не меньшее k.

Если в интервал попадает менее 5 данных, то этот интервал (для корректности применения некоторых вычислительных процедур) иногда следует объединить с другим интервалом или интервалами, чтобы количество попавших в них данных было 5 или более.

Если числовое значение данных выборки совпадает со значением границ двух расположенных рядом интервалов, то такие данные можно поместить в любой из двух интервалов произвольно.

Точечные оценки параметров статистических распределений

Точечной называют оценку параметра, которая определяется одним числом.

Генеральной средней $\bar{x}_{\text{г}}$ называется среднее взвешенное всех значений генеральной совокупности, определяемое по формуле $\bar{x}_{\text{г}} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i w_i$, где k – количество вариант (интервалов) статистического распределения.

Выборочной средней $\bar{x}_{\text{в}}$ называется среднее взвешенное всех значений выборки, определяемое по формуле $\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i w_i$, где k – количество вариант (интервалов) статистического распределения.

Генеральной дисперсией $D_{\text{г}}$ называется среднее арифметическое квадратов отклонений значений генеральной совокупности от генеральной средней, определяемое по формуле $D_{\text{г}} = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x}_{\text{г}})^2 n_i = \sum_{i=1}^{k}(x_i - \bar{x}_{\text{г}})^2 w_i = \sum_{i=1}^{k} x_i^2 w_i - \left(\bar{x}_{\text{г}}\right)^2$.

Выборочной дисперсий $D_{\text{в}}$ называется среднее арифметическое квадратов отклонений значений выборки от выборочной средней, определяемое по формуле $D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x}_{\text{в}})^2 n_i = \sum_{i=1}^{k}(x_i - \bar{x}_{\text{в}})^2 w_i = \sum_{i=1}^{k} x_i^2 w_i - \left(\bar{x}_{\text{в}}\right)^2$.

Генеральным средним квадратическим отклонением $\sigma_{\text{г}}$ называют квадратный корень из генеральной дисперсии: $\sigma_{\text{г}} = \sqrt{D_{\text{г}}}$.

Выборочным средним квадратическим отклонением (стандартной ошибкой выборки) $\sigma_{\text{в}}$ называют квадратный корень из выборочной дисперсии: $\sigma_{\text{в}} = \sqrt{D_{\text{в}}}$.

Модой M_0* вариационного ряда называется варианта, имеющая наибольшую частоту.

Медианой M^*_e вариационного ряда называется значение, приходящееся на середину ряда $x_1, x_2, \ldots, x_k$. Если $k = 2l + 1$, то $M^*_e = x_{l+1}$. Если $k = 2l$, то $M_e^* = \frac{x_l + x_{l+1}}{2}$.

Начальный эмпирический момент M_s *порядка* s статистического распределения определяют по формуле $M_s = \sum_{i=1}^{k} x_i^s w_i$, где x_i – наблюдаемое значение признака, w_i – относительная частота наблюдаемого значения признака. Начальный эмпирический момент первого порядка равен генеральной или выборочной средней $M_1 = \bar{x}_{\text{в}}$ или $M_1 = \bar{x}_{\text{г}}$.

Центральный эмпирический момент m_s *порядка* s статистического распределения определяют по формуле $m_s = \sum_{i=1}^{k} \left(x_i - \bar{x}_{\text{г}}\right)^s w_i$ или $m_s = \sum_{i=1}^{k} \left(x_i - \bar{x}_{\text{в}}\right)^s w_i$. Центральный эмпирический момент второго порядка равен генеральной или выборочной дисперсии $m_2 = D_{\text{г}}$ или $m_2 = D_{\text{в}}$.

Центральные и начальные эмпирические моменты связаны следующими соотношениями:

$$m_2 = M_2 - M_1^2;\ m_3 = M_3 - 3M_2M_1 + 2M_1^3;$$
$$m_4 = M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4.$$

Коэффициент асимметрии A_s^* статистического распределения определяется по формуле $A_s^* = \frac{m_3}{\sigma_{\text{г}}^3} = \frac{1}{\sigma_{\text{г}}^3} \sum_{i=1}^{k} \left(x_i - \bar{x}_{\text{г}}\right)^3 w_i$ или $A_s^* = \frac{m_3}{\sigma_{\text{в}}^3} = \frac{1}{\sigma_{\text{в}}^3} \sum_{i=1}^{k} \left(x_i - \bar{x}_{\text{в}}\right)^3 w_i$.

Эксцесс E_x^* статистического распределения определяется по формуле $E_x^* = \frac{m_4}{\sigma_{\text{г}}^4} - 3 = \frac{1}{\sigma_{\text{г}}^4} \sum_{i=1}^{k} \left(x_i - \bar{x}_{\text{г}}\right)^4 w_i - 3$ или $E_x^* = \frac{m_4}{\sigma_{\text{в}}^4} - 3 = \frac{1}{\sigma_{\text{в}}^4} \sum_{i=1}^{k} \left(x_i - \bar{x}_{\text{в}}\right)^4 w_i - 3$.

Относительной характеристикой рассеивания случайной величины выступает *коэффициент вариации* V, который вычисляется как отношение среднего квадратического отклонения к средней по формуле $V = \frac{\sigma_{\text{г}}}{\bar{x}_{\text{г}}}$ или $V = \frac{\sigma_{\text{в}}}{\bar{x}_{\text{в}}}$.

Состоятельной называют статистическую оценку параметра, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру. Если дисперсия несмещенной оценки при $n \to \infty$ стремится к нулю, то такая оценка оказывается и состоятельной.

Оценка θ^* параметра θ называется *несмещенной*, если ее математическое ожидание равно оцениваемому параметру при любом объеме выборки $M(\theta^*) = \theta$, в противном случае оценка называется *смещенной*.

Выборочная средняя $\overline{x}_{\text{в}}$ – *несмещенная оценка математического ожидания* случайной величины *X*.

Выборочная дисперсия $D_{\text{в}}$ является *смещенной* оценкой дисперсии случайной величины *X*. *Несмещенной оценкой дисперсии* случайной величины *X* выступает исправленная дисперсия $s_{\text{в}}^2 = \dfrac{n}{n-1} \cdot D_{\text{в}}$.

Эффективной называют статистическую оценку, которая при заданном объеме выборки *n* имеет наименьшую возможную дисперсию.

Пример. Три выборочные совокупности (см ниже) были извлечены из трех независимых генеральных совокупностей. Необходимо составить статистические распределения выборок и определить их числовые характеристики.

Выборка 1

17,17	13,88	16,92	30,84	9,78	8,12	26,15	20,69	31,67	24,12
10,42	9,09	16,53	9,28	12,76	30,23	23,04	27,12	30,21	25,56
22,32	8,78	29,85	24,92	9,54	10,41	12,17	27,34	29,69	22,03
29,58	11,94	19,18	27,60	16,60	14,16	17,72	14,29	21,08	11,65
29,23	13,27	18,23	31,34	19,69	26,62	21,26	12,27	20,02	29,41
31,00	8,41	15,29	19,19	20,27	24,31	25,08	28,80	24,20	17,07
8,35	14,84	31,42	15,21	16,96	27,42	21,32	10,76	19,76	12,81
17,78	16,23	27,36	26,00	31,66	25,38	12,35	9,43	11,50	12,94
28,72	21,29	31,79	16,44	8,98	10,04	31,29	26,28	8,91	16,02
11,33	16,58	14,15	26,62	13,54	11,17	24,49	25,72	27,11	15,80

Выборка 2

7,1	0,96	22,79	8,36	3	3,6	7,94	8,78	7,31	10,87
2,29	1,08	4,8	37,81	3,12	3,72	8,15	8,99	9,83	16,77
5,63	1,2	4,92	6,89	3,24	3,84	12,03	9,2	11,16	17,36
0,48	1,32	9,62	0,84	35,98	3,96	32,72	10,29	11,45	20,93
0,6	4,08	2,28	1,56	3,48	14,93	24,65	10,58	11,74	21,86
0,72	4,2	2,4	18,54	6,68	1,8	6,26	13,48	5,21	3,36
12,32	4,32	2,52	7,52	14,35	1,92	6,47	13,77	5,42	4,56
6,05	4,44	2,64	9,41	14,64	2,04	19,13	0,12	15,59	26,83
1,68	39,64	2,76	5,84	12,61	2,16	0,24	17,95	16,18	34,15
1,44	4,68	2,88	7,73	12,9	14,06	19,72	0,36	13,19	8,57

Выборка 3

14,22	13,14	16,78	12,64	11,91	14,43	19,11	19,58	19,03	19,01
12,33	13,01	16,23	8,72	6,59	16,81	13,58	9,87	19,96	12,80
14,09	20,72	16,66	15,19	17,04	11,64	20,02	11,00	10,07	12,13
23,74	13,54	20,76	23,38	16,01	21,11	11,95	17,06	21,77	12,56
8,68	16,59	12,56	15,63	23,46	22,60	22,85	15,03	9,42	22,54
3,18	17,64	17,98	20,70	15,58	10,12	11,11	13,98	12,30	19,58
19,05	22,25	18,64	13,81	21,38	14,30	7,93	14,87	24,25	13,49
16,03	17,79	5,05	19,58	24,74	11,74	21,74	15,33	12,60	8,79
26,38	16,93	11,00	7,52	18,55	16,16	20,05	14,18	23,71	16,18
17,83	14,63	15,38	14,63	20,84	19,74	23,40	16,12	13,15	12,86

Решение

Характеристики выборок 1 – 3 представлены далее в таблице. Максимальные $X_{\max}$ и минимальные $X_{\min}$ значения выборок находятся с помощью функций Excel «МАКС» и «МИН» из категории «Статистические». Определение количества интервалов для составления интервального статистического распределения проводится по формуле Старджесса: $k = 1 + 3{,}322 \lg 100 = 7{,}644 \approx 8$.

Характеристики выборок

№ выборки	$X_{\min}$	$X_{\max}$	R	k	h
1	$8{,}12 \approx 8$	$31{,}79 \approx 32$	24	8	3
2	$0{,}12 \approx 0$	$39{,}64 \approx 40$	40	8	5
3	$3{,}18 \approx 3$	$26{,}38 \approx 27$	24	8	3

В ячейки электронного процессора Excel вводим значения нижней границы первого интервала и значения верхних границ всех интервалов. Используя подпрограмму «Гистограмма» (*Сервис → Надстройки → Анализ данных → Гистограмма*), получим первоначальные статистические распределения выборок (см. следующую таблицу)

№ интервала	Выборка 1			Выборка 2			Выборка 3		
	Границы интервала	Среднее	Частота	Границы интервала	Среднее	Частота	Границы интервала	Среднее	Частота
1	(8; 11)	9,5	15	(0; 5)	2,5	42	(3; 6)	4,5	2
2	(11; 14)	12,5	14	(5; 10)	7,5	23	(6; 9)	7,5	6
3	(14; 17)	15,5	15	(10; 15)	12,5	17	(9; 12)	10,5	11

4	(17; 20)	18,5	9	(15; 20)	17,5	8	(12; 15)	13,5	25
5	(20; 23)	21,5	9	(20; 25)	22,5	4	(15; 18)	16,5	23
6	(23; 26)	24,5	11	(25; 30)	27,5	1	(18; 21)	19,5	17
7	(26; 29)	27,5	12	(30; 35)	32,5	2	(21; 24)	22,5	13
8	(29; 32)	30,5	15	(35; 40)	37,5	3	(24; 27)	25,5	3
		Еще	0		Еще	0		Еще	0
		Сумма	100		Сумма	100		Сумма	100

В каждом из интервалов №№ 5–8 выборки 2 и в интервалах №№ 1, 8 выборки 3 наблюдается менее 5 чисел, следовательно, необходимо произвести объединение интервалов, но такое преобразование приведет к тому, что интервальные статистические распределения 2-й и 3-й выборок будут иметь менее 7 интервалов (ниже допустимого минимума). В этом случае необходимо уменьшить длину интервалов, увеличивая при этом их количество. Значение размаха $R = 40$ выборки 2 при делении на 10 дает число 4, значит целесообразно принять его за новую длину интервала. Значение размаха $R = 24$ выборки 3 при делении на 12 дает число 2, поэтому примем его за новую длину интервала, как показано в следующей таблице.

№ интервала	Выборка 2			Выборка 3		
	Границы интервала	Среднее	Частота	Границы интервала	Среднее	Частота
1	(0; 4)	2	34	(3; 5)	4	1
2	(4; 8)	6	22	(5; 7)	6	2
3	(8; 12)	10	15	(7; 9)	8	5
4	(12; 16)	14	12	(9; 11)	10	6
5	(16; 20)	18	7	(11; 13)	12	14
6	(20; 24)	22	3	(13; 15)	14	16
7	(24; 28)	26	2	(15; 17)	16	17
8	(28; 32)	30	0	(17; 19)	18	8
9	(32; 36)	34	3	(19; 21)	20	15
10	(36; 40)	38	2	(21; 23)	22	8
11		Еще	0	(23; 25)	24	7
12		Сумма	100	(25; 27)	26	1
					Еще	0
					Сумма	100

Объединение некоторых интервалов (№№ 6–10 в выборке 2 №№ 1–3, 11–12 в выборке 3) приведет к тому, что их статистические распределения будут иметь 7 и 9 интервалов соответственно. В следующей таблице вместо частот показаны результаты их деления на ширину соответствующих интервалов, что является аналогом плотности распределения случайной величины.

	Выборка 1		Выборка 2		Выборка 3	
i	Границы интервала	$\frac{w_i}{h}$	Границы интервала	$\frac{w_i}{h}$	Границы интервала	$\frac{w_i}{h}$
1	(8; 11)	0,05	(0; 4)	0,085	(3; 9)	0,013
2	(11; 14)	0,047	(4; 8)	0,055	(9; 11)	0,03
3	(14; 17)	0,05	(8; 12)	0,038	(11; 13)	0,07
4	(17; 20)	0,03	(12; 16)	0,03	(13; 15)	0,08
5	(20; 23)	0,03	(16; 20)	0,018	(15; 17)	0,085
6	(23; 26)	0,037	(20; 28)	0,006	(17; 19)	0,04
7	(26; 29)	0,04	(28; 40)	0,004	(19; 21)	0,075
8	(29; 32)	0,05			(21; 23)	0,04
9					(23; 27)	0,02

По данным этой таблицы можно построить гистограммы, которые наглядно иллюстрируют выборки (см. рисунки справа).

Вычисление характеристик выборки 1:

$$\bar{x}_{\text{в}} = \frac{1}{100}(9{,}5 \cdot 15 + 12{,}5 \cdot 14 + \cdots) = 19{,}7$$

$$D_{\text{в}} = \frac{1}{100}(9{,}5^2 \cdot 15 + 12{,}5^2 \cdot 14 + \cdots) - 19{,}7^2 = 52{,}1$$

$$\sigma_{\text{в}} = \sqrt{52{,}1} = 7{,}2$$

Вычисление среднего для выборки 2:

$$\bar{x}_{\text{в}} = \frac{1}{100}(2 \cdot 34 + 6 \cdot 22 + \cdots) = 9{,}34$$

Вычисление характеристик выборки 3:

$$\bar{x}_{\text{в}} = \frac{1}{100}(6 \cdot 8 + 10 \cdot 6 + 12 \cdot 14 \ldots) \approx 16$$

$$D_{\text{в}} = \frac{1}{100}(6^2 \cdot 8 + 10^2 \cdot 6 + 12^2 \cdot 14 \ldots) - 16^2 \approx 23$$

$$\sigma_{\text{в}} = \sqrt{23} \approx 5$$

15.2. Проверка статистических гипотез о виде распределения

Исследование видов распределений генеральных совокупностей на основе выборок включает два этапа: 1) составление статистических распределений выборочных совокупностей; 2) определение законов статистических распределений генеральной совокупности.

Статистические гипотезы

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений. Статистические гипотезы бывают двух видов: нулевая (выдвигаемая) гипотеза H_0 (о том, как формулировать правдоподобные гипотезы, будет сказано далее) и конкурирующая (противоречащая нулевой) H_1.

Проведение проверки статистических гипотез статистическими методами приводит к появлению ошибок двух родов: *ошибка первого рода* – отвержение правильной гипотезы; *ошибка второго рода* – принятие неправильной гипотезы.

Вероятность совершить ошибку первого рода называют *уровнем значимости* и обозначают через α (вероятность отвергнуть правильную гипотезу). Наиболее часто уровень значимости принимают 0,05, что означает наличие риска отвергнуть правильную гипотезу в пяти случаях из ста.

Для проверки нулевой гипотезы H_0 используется специально подобранная случайная величина, которая называется *статистическим критерием К*. *Наблюдаемым значением критерия* $K_{\text{набл}}$ называют его значение, вычисленное по выборке. После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза H_0 отвергается, а другое – при которых она принимается.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу H_0 отвергают.

Областью принятия гипотезы называют совокупность значений критерия, при которых нулевую гипотезу H_0 принимают.

Критической точкой $K_{\text{кр}}$ называют значение критерия *К*, которое отделяет критическую область от области принятия гипотезы. Для каждого

критерия имеются соответствующие таблицы, по которым и находят критическую точку.

Правосторонней называют критическую область, определяемую неравенством $K > K_{кр}$ ($K_{кр}>0$). Критические точки $K_{кр}$ правосторонней критической области находят из условия $P(K > K_{кр})=\alpha$, где α – достаточно малая вероятность отвергнуть правильную гипотезу (малое значение уровня значимости α). В этом случае вероятность *не отвергнуть* правильную гипотезу $(1-\alpha)$ будет равна значению функции распределения $F(K)=P(K<K_{кр})=1-\alpha$ случайной величины К.

Левосторонней называют критическую область, определяемую неравенством $K<K_{кр}$ ($K_{кр}<0$). Критические точки $K_{кр}$ левосторонней критической области находят из условия $P(K < K_{кр})=\alpha$.

Двусторонней называют критическую область, определяемую неравенствами $K<K_{кр1}$, $K>K_{кр2}$, $K_{кр1}<0$, $K_{кр2}>0$ Критические точки $K_{кр}$ двусторонней критической области находят из условия $P(K>K_{кр1})+P(K>K_{кр2})=\alpha$, для симметричной критической области $P(K > |K_{кр}|) = \alpha/2$.

Критические точки $K_{кр}$ двусторонней и правосторонней критических областей соответствуют квантилям распределения случайной величины *K*.

Основной принцип проверки статистических гипотез для правосторонней и двусторонней областей: если наблюдаемое значение критерия $K_{набл}$ меньше критической точки $K_{кр}$ или $F(K_{набл}) \in (0; F(K_{кр}))$, то нулевую гипотезу H_0 принимают, если наблюдаемое значение критерия $K_{набл}$ больше критической точки $K_{кр}$ или $F(K_{набл}) \notin (0; F(K_{кр}))$, то нулевую гипотезу H_0 отклоняют.

Определение законов статистических распределений (формулировка гипотез)

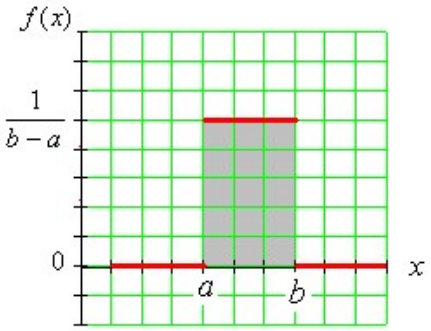
Формально, процедуры верификации (подтверждения) или фальсификации (отвержения) гипотез могут применяться к разным утверждениям о свойствах выборки, однако полезно попытаться изначально сформулировать такие утверждения, которые могут быть проверены и имеют шансы на положительный результат проверки. Основой для выдвижения гипотезы о виде закономерности выборочной совокупности выступает совпадение формы фигуры, ограниченной графиком $f(x)$ плотности

непрерывной случайной величины X и осью Ox, с гистограммой плотности относительных частот статистического распределения выборки.

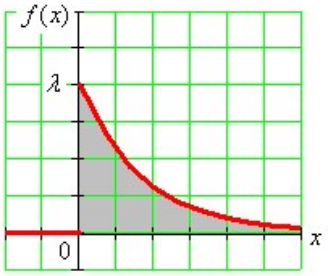
Пример. Для рассмотренных ранее выборок 1 – 3 выдвинуть предварительные предположения о виде распределения.

Решение

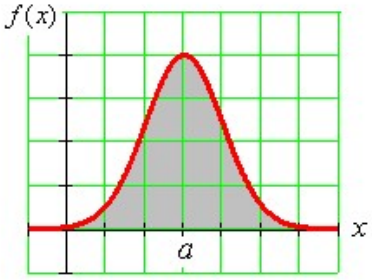
Вернемся к полученным ранее гистограммам и сравним их с типовым видом графиков случайных величин, имеющих равномерное, показательное и нормальное распределения



равномерное



показательное



нормальное

Внешнее приблизительное сходство закрашенных фигур на гистограммах и графиках позволяет выдвинуть следующие нулевые гипотезы:

H_0^1: генеральная совокупность *первая*, из которой извлечена выборка 1, имеет равномерное распределение при значимости $\alpha = 0{,}05$.

H_0^2: генеральная совокупность *вторая*, из которой извлечена выборка 2, имеет экспоненциальное распределение при уровне значимости $\alpha = 0{,}05$.

H_0^3: генеральная совокупность *третья*, из которой извлечена выборка 3, имеет нормальное распределение при уровне значимости $\alpha = 0{,}05$.

Метод моментов для установления параметров гипотететических распределений

Дальнейшее установление закона распределения выборочной совокупности проводится через проверку нулевой H_o статистической гипотезы о виде предполагаемого распределения непрерывной случайной величины X с параметрами, определяемыми методом моментов, при заданном уровне значимости α для числа степеней свободы r.

Метод моментов – это определение неизвестных параметров статистического распределения путем приравнивания моментов случайной непрерывной величины соответствующим моментам того же порядка статистического распределения. Основой рассматриваемого подхода выступает тот факт, что начальные и центральные моменты статистического распределения являются состоятельными оценками соответственно начальных и центральных моментов того же порядка непрерывной случайной величины X. Метод моментов предложен К. Пирсоном.

Параметры a и b равномерного распределения можно найти, если приравнять:

1. Начальный момент первого порядка равномерного распределения случайной величины X к начальному эмпирическому моменту M_1 первого порядка статистического распределения $\nu_1 = M_1$ или $\bar{x}_в = \frac{a+b}{2}$;

2. Центральный момент μ_2 второго порядка равномерного распределения случайной величины X к центральному эмпирическому моменту m_2 второго порядка статистического распределения $\mu_2 = m_2$ или $D_в = \frac{(b-a)^2}{12}$.

Значит, параметры равномерного распределения a и b вычисляются по формулам: $a = \bar{x}_в - \sqrt{3}\sigma_в$; $b = \bar{x}_в + \sqrt{3}\sigma_в$.

Параметр λ экспоненциального распределения можно найти, если приравнять начальный момент ν_1 первого порядка показательного распределения случайной величины X начальному эмпирическому моменту M_1 первого порядка статистического распределения $\nu_1 = M_1$ или $\bar{x}_в = \frac{1}{\lambda}$.

Параметры a и σ *нормального распределения* можно найти, если приравнять:

1. Начальный момент первого ν_1 порядка нормального распределения случайной величины X к начальному эмпирическому моменту M_1 первого порядка статистического распределения $\nu_1 = M_1$ или $\bar{x}_в = a$.

2. Центральный момент μ_2 второго порядка нормального распределения случайной величины X к центральному эмпирическому моменту m_2 второго порядка статистического распределения $\mu_2 = m_2$ или $\sigma = \sigma_в$.

Начальные и центральные эмпирические моменты третьего и четвертого порядков статистического распределения приравниваются соответственно к начальным и центральным моментам третьего и четвертого порядков случайной величины X.

Процедура проверки гипотез

Статистическим критерием проверки гипотезы о закономерности распределения выборочной совокупности выступает критерий Пирсона χ^2 (хи-квадрат). Вычисление наблюдаемого значения $\chi^2_{набл}$ критерия Пирсона связано с нахождением для каждого интервала случайной величины $\frac{(n_i - n_i')^2}{n_i'}$, где n_i и n_i' – это соответственно частоты статистического распределения и распределения непрерывной случайной величины X: $\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$, где k – количество интервалов.

Частота n_i статистического распределения равна количеству случайных величин выборки, попавших в данный интервал. Частота n_i' непрерывной

случайной величины X вычисляется по формуле $n_i^{'} = n \cdot P_i$, где P_i – вероятность попадания непрерывной случайной величины X в интервал (x_i, x_{i+1}), n – объем выборки.

Поскольку односторонний критерий более жестко отвергает нулевую гипотезу, чем двусторонний, для проверки гипотезы целесообразно выбрать правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α: $P\left(\chi^2_{\text{набл}} > \chi^2_{\text{кр}}(\alpha, r)\right) = \alpha$. Критические точки $\chi^2_{\text{кр}}(\alpha, r)$ распределения Пирсона находят, используя функцию Excel ХИ2ОБР (вероятность α; степени свободы r).

Нулевую гипотезу при заданном уровне значимости α для числа степеней свободы r следует принять, если наблюдаемое значение критерия Пирсона меньше значения критической точки $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}(\alpha, r)$. Нулевую гипотезу следует отвергнуть, если наблюдаемое значение критерия Пирсона больше значения критической точки $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}(\alpha, r)$.

Число степеней свободы r определяется по формуле $r = k - 1 - g$, где g – это количество параметров предполагаемого распределения (для нормального и равномерного распределений $g = 2$, для экспоненциального распределения $g = 1$).

Можно для подтверждения выдвигаемой гипотезы сравнивать полученные значения некоторых точечных оценок со значениями постоянных величин этих оценок известных распределений непрерывной случайной величины: коэффициент асимметрии A_s^* статистического распределения с коэффициентами асимметрии равномерного и нормального распределений $A_s = 0$; эксцесс E_x^* статистического распределения с эксцессами равномерного ($E_x = -1{,}2$) или нормального ($E_x = 0$) распределений; коэффициент вариации V статистического распределения с коэффициентами вариации нормального ($V \le 0{,}3$) или показательного ($V = 1$) распределений или использовать эти факты при выборе вида распределения.

Пример (проверка гипотезы о равномерном распределении). Проверить статистическую гипотезу H_0^1: генеральная совокупность *первая*, из которой извлечена выборка 1, имеет равномерное распределение при уровне значимости $\alpha = 0{,}05$.

Решение

Условию построения правосторонней критической области при уровне значимости $\alpha = 0{,}05$ соответствует равенство $P\left(\chi^2_{\text{набл}} > \chi^2_{\text{кр}}(0{,}05;\ r)\right) = 0{,}05.$

Вычисление значения критической точки $\chi^2_{\text{кр}}$ распределения Пирсона (хи-квадрат) при уровне значимости $\alpha = 0{,}05$ для числа степеней свободы $r = 8 - 2 - 1 = 5$ осуществляется с помощью формулы Excel ХИ2ОБР (α; r): у нас $\chi^2_{\text{кр}}(0{,}05;\ 5)$=ХИ2ОБР(0,05;5)=11,07 и условие принятия статистической гипотезы H_0^1 оказывается известным заранее: $\chi^2_{\text{набл}} < 11{,}07.$

Теоретические частоты предполагаемого равномерного распределения определяются по ниже приведенным формулам:

$$n_1' = n \cdot \frac{1}{b-a}(x_1 - a); \quad x_1 - \text{верхняя граница } 1-\text{го интервала};$$

$$n_2' = n_3' = \ldots = n_{k-1}' = n \cdot \frac{1}{b-a}(x_i - x_{i-1}) = n \cdot \frac{1}{b-a} \cdot h, \quad i = 2,\ldots,k-1;$$

$$n_k' = n \cdot \frac{1}{b-a}(b - x_{k-1}); \; x_{k-1} - \text{нижняя граница } k-\text{го интервала}.$$

Параметры предполагаемого равномерного распределения вычисляются по формулам: $a = \bar{x}_{\text{в}} - \sqrt{3}\sigma_{\text{в}};\ b = \bar{x}_{\text{в}} + \sqrt{3}\sigma_{\text{в}}$. Точечные оценки параметров статистического распределения выборки 1 были найдены ранее: $\bar{x}_{\text{в}} = 19{,}7;$ $\sigma_{\text{в}} = 7{,}2.$ Вычисление параметров предполагаемого равномерного распределения: $a = 19{,}7 - 7{,}2 \cdot \sqrt{3} = 7{,}2$ и $b = 19{,}7 + 7{,}2 \cdot \sqrt{3} = 32{,}2$.

Вычисление частот предполагаемого равномерного распределения:

$$n_1' = 100 \cdot \frac{1}{32{,}2 - 7{,}2} \cdot (11 - 7{,}2) = 15{,}2;$$

$$n_2' = n_3' = n_4' = n_5' = n_6' = n_7' = 100 \cdot \frac{1}{32{,}2 - 7{,}2} \cdot 3 = 12;$$

$$n_8' = 100 \cdot \frac{1}{32{,}2 - 7{,}2} \cdot (32{,}2 - 29) = 12{,}8.$$

Вычисление значения $\chi^2_{\text{набл}}$для выборки 1 представлено в следующей таблице:

i	x_{i-1}	x_i	n_i	n_i'	$\frac{(n_i - n_i')^2}{n_i'}$
1	8	11	15	15,2	0,003
2	11	14	14	12	0,333
3	14	17	15	12	0,750
4	17	20	9	12	0,750
5	20	23	9	12	0,750
6	23	26	11	12	0,083
7	26	29	12	12	0,000
8	29	32	15	12,8	0,378
				$\sum$	3,05

Значение $\chi^2_{\text{набл}} = 3{,}05$ меньше значения критической точки $\chi^2_{\text{кр}} = 11{,}07$, следовательно, нет оснований отвергнуть гипотезу $H_0{}^1$ о равномерном распределении генеральной совокупности *первой* при уровне значимости $\alpha = 0{,}05$.

Пример (проверка гипотезы о показательном распределении). Проверить статистическую гипотезу $H_0{}^2$: генеральная совокупность *вторая*, из которой извлечена выборка 2, имеет показательное (экспоненциальное) распределение при уровне значимости $\alpha = 0{,}05$.

Решение

Условию построения правосторонней критической области при уровне значимости $\alpha = 0{,}05$ соответствует равенство $P\left(\chi^2_{\text{набл}} > \chi^2_{\text{кр}}(0{,}05;\ r)\right) = 0{,}05.$

Вычисление значения критической точки $\chi^2_{\text{кр}}$ распределения Пирсона (хи-квадрат) при уровне значимости $\alpha = 0{,}05$ для числа степеней свободы $r = 7 - 1 - 1 = 5$ осуществляется с помощью формулы Excel ХИ2ОБР (α; r): у нас $\chi^2_{\text{кр}}(0{,}05;\ 5)$=ХИ2ОБР(0,05;5)=11,07 и условие принятия статистической гипотезы $H_0{}^1$ оказывается известным заранее: $\chi^2_{\text{набл}} < 11{,}07.$

Параметр предполагаемого экспоненциального распределения вычисляется по формуле $\lambda = \frac{1}{\bar{x}_в} = \frac{1}{9{,}34} \approx 0{,}1$ (среднее выборочное для выборки 2 было найдено ранее). Частоты предполагаемого экспоненциального распределения вычисляются по формуле $n_i' = n\left(e^{-\lambda x_{i-1}} - e^{-\lambda x_i}\right)$. Иначе, используя универсальный подход и автоматизированные вычисления MS Excel, можно поступить так: $n_i' = nP_i = n\left(F(x_i) - F(x_{i-1})\right)$, где:

$F(0)$ = ЭКСПРАСП (0; 0,1; ИСТИНА) = 0,000;

$F(4)$ = ЭКСПРАСП (4; 0,1; ИСТИНА) = 0,330;

$F(8)$ = ЭКСПРАСП (8; 0,1; ИСТИНА) = 0,551;

$F(12)$ = ЭКСПРАСП (12; 0,1; ИСТИНА) = 0,699;

$F(16)$ = ЭКСПРАСП (16; 0,1; ИСТИНА) = 0,798;

$F(20)$ = ЭКСПРАСП (20; 0,1; ИСТИНА) = 0,865;

$F(28)$ = ЭКСПРАСП (28; 0,1; ИСТИНА) = 0,939;

$F(40)$ = ЭКСПРАСП (40; 0,1; ИСТИНА) = 0,982.

Вычисление значения $\chi^2_{\text{набл}}$для выборки 2 представлено в следующей таблице:

i	x_i	x_{i+1}	n_i	$F(x_i)$	$F(x_{i+1})$	P_i	$n_i' = P_i \cdot n$	$\frac{(n_i - n_i')^2}{n_i'}$
1	0	4	34	0,000	0,330	0,330	33	0,0303
2	4	8	22	0,330	0,551	0,221	22,1	0,0005
3	8	12	15	0,551	0,699	0,148	14,8	0,0027
4	12	16	12	0,699	0,798	0,099	9,9	0,4455
5	16	20	7	0,798	0,865	0,067	6,7	0,0134
6	20	28	5	0,865	0,939	0,074	7,4	0,7784
7	28	40	5	0,939	0,982	0,043	4,3	0,1140
							$\sum$	1,38

Значение $\chi^2_{\text{набл}} = 1{,}38$ меньше значения критической точки $\chi^2_{\text{кр}} = 11{,}07$, следовательно, нет оснований отвергнуть гипотезу $H_0{}^2$ об экспоненциальном распределении генеральной совокупности *второй* при уровне значимости $\alpha = 0{,}05$.

Пример (проверка гипотезы о нормальном распределении). Проверить статистическую гипотезу H_0^3: генеральная совокупность *третья*, из которой извлечена выборка 3, имеет нормальное распределение при уровне значимости $\alpha = 0{,}05$.

Решение

Условию построения правосторонней критической области при уровне значимости $\alpha = 0{,}05$ соответствует равенство $P\left(\chi^2_{\text{набл}} > \chi^2_{\text{кр}}(0{,}05;\ r)\right) = 0{,}05.$

Вычисление значения критической точки $\chi^2_{\text{кр}}$ распределения Пирсона (хи-квадрат) при уровне значимости $\alpha = 0{,}05$ для числа степеней свободы $r = 9 - 2 - 1 = 6$ осуществляется с помощью формулы Excel ХИ2ОБР (α; r): у нас $\chi^2_{\text{кр}}(0{,}05;\ 6)$=ХИ2ОБР(0,05;6)=12,59 и условие принятия статистической гипотезы H_0^1 оказывается известным заранее: $\chi^2_{\text{набл}} < 12{,}59.$

Точечные оценки математического ожидания (среднее выборочное) и среднеквадратического отклонения (по выборочному аналогу) для выборки 3 были получены ранее: $\bar{x}_{\text{в}} \approx 16;\ \sigma_{\text{в}} \approx 5.$

Частоты предполагаемого нормального распределения вычисляются по формуле $n_i' = nP_i = n\big(F(x_i) - F(x_{i-1})\big)$, где $F(x)$ – функция нормального распределения. Используя возможности MS Excel, получаем:

F(3) = НОРМРАСП (3; 16; 5; ИСТИНА) = 0,005;

F(9) = НОРМРАСП (9; 16; 5; ИСТИНА) = 0,081;

F(11) = НОРМРАСП (11; 16; 5; ИСТИНА) = 0,159;

F(13) = НОРМРАСП (13; 16; 5; ИСТИНА) = 0,274;

F(15) = НОРМРАСП (15; 16; 5; ИСТИНА) = 0,421;

F(17) = НОРМРАСП (17; 16; 5; ИСТИНА) = 0,579;

F(19) = НОРМРАСП (19; 16; 5; ИСТИНА) = 0,726;

F(21) = НОРМРАСП (21; 16; 5; ИСТИНА) = 0,841;

F(23) = НОРМРАСП (23; 16; 5; ИСТИНА) = 0,919.

F(27) = НОРМРАСП (27; 16; 5; ИСТИНА) = 0,986.

Вычисление значения $\chi^2_{\text{набл}}$для выборки 3 представлено в следующей таблице:

i	x_i	x_{i+1}	n_i	$F(x_i)$	$F(x_{i+1})$	P_i	$n_i^{'} = P_i \cdot n$	$\frac{(n_i - n_i^{'})^2}{n_i^{'}}$
1	3	9	8	0,005	0,081	0,076	7,6	0,021
2	9	11	6	0,081	0,159	0,078	7,8	0,415
3	11	13	14	0,159	0,274	0,115	11,5	0,543
4	13	15	16	0,274	0,421	0,147	14,7	0,115
5	15	17	17	0,421	0,579	0,158	15,8	0,091
6	17	19	8	0,579	0,726	0,147	14,7	3,054
7	19	21	15	0,726	0,841	0,115	11,5	1,065
8	21	23	8	0,841	0,919	0,078	7,8	0,005
9	23	27	8	0,919	0,986	0,067	6,7	0,252
							$\sum$	5,56

Значение $\chi^2_{\text{набл}} = 5{,}56$ меньше значения критической точки $\chi^2_{\text{кр}} = 12{,}59$, следовательно, нет оснований отвергнуть гипотезу H_0^3 о нормальном распределении генеральной совокупности *третьей* при уровне значимости $\alpha = 0{,}05$.