Проверка гипотезы о нормальном распределении

Ранее мы обсудили свойства нормального распределения. Рассмотрим теперь весьма важную практическую проблему. Насколько естественным является предположение о том, что конкретные данные представляют собой значения нормально распределенной случайной величины?[1] Для ответа на этот вопрос используется один из следующих исследовательских методов:

1. Сравнение характеристик набора данных со свойствами нормального распределения.
2. Построение специального графика на основе набора данных.

Оценка свойств

Напомним, что нормальное распределение является симметричным и колоколообразным, так что все характеристики его среднего значения — математическое ожидание, мода и медиана — совпадают друг с другом. Межквартильный размах нормального распределения равен 1,33 стандартного отклонения. Нормальное распределение является непрерывным, причем нормально распределенная случайная величина принимает произвольные значения, лежащие на всей числовой оси.

На практике характеристики набора данных могут немного отличаться от теоретических, либо потому, что случайная величина является лишь приближенно нормальной, либо потому, что ее реальные свойства отличаются от предполагаемых. В таких ситуациях кривая распределения оказывается не совсем симметричной и колоколообразной. Оценки математического ожидания могут слегка отличаться от теоретических, а межквартильный размах может не быть равным 1,33 стандартного отклонения. Кроме того, на практике диапазон изменения данных не может быть бесконечным — как правило, он ограничен шестью стандартными отклонениями. Такие распределения являются приближенно нормальными.

Многие непрерывные случайные величины не являются ни точно, ни приближенно нормальными. Свойства таких величин довольно сильно отличаются от свойств нормального распределения, перечисленных выше. Рассмотрим, например, оценки, полученные студентами при сдаче четырех тестов (рис. 1). Excel справляется с обработкой данных, не требуя их упорядочения. Вычислим описательные статистики результатов каждого теста в отдельности с помощью надстройки *Анализ данных* (как это сделать, см., например, Представление числовых данных в виде таблиц и диаграмм).

	A	B	C	D	E	F	G	H	I	J
1	Тест 1	Тест 2	Тест 3	Тест 4						
2	70	77	64	77			*Тест 1*	*Тест 2*	*Тест 3*	*Тест 4*
3	69	77	62	74						
4	64	73	55	62		Среднее (математическое ожидание)	65,0	70,7	59,3	65,0
5	55	58	50	44		Стандартная ошибка	2,1	2,3	2,3	3,9
6	72	78	66	80		Медиана	65,0	74,0	56,0	65,0
7	68	76	59	71		~~Мода~~	~~#Н/Д~~	~~77,0~~	~~53,0~~	~~#Н/Д~~
8	52	54	48	41		Стандартное отклонение	9,0	10,0	10,0	16,9
9	60	66	53	53		Дисперсия выборки	80,2	100,0	100,0	285,0
10	62	71	54	59		Эксцесс	-0,4	0,2	0,2	-1,2
11	58	64	52	50		Асимметричность	0,0	-1,0	1,0	0,0
12	78	82	76	89		Интервал (размах)	34	36	36	54
13	82	83	83	92		Минимум	48	47	47	38
14	66	75	57	68		Максимум	82	83	83	92
15	65	74	56	65		Сумма	1235	1343	1127	1235
16	57	61	51	47		Счет (объем выборки)	19	19	19	19
17	75	80	72	86						
18	48	47	47	38		Межквартильный размах	12,0	12,5	12,5	27,0
19	61	68	53	56		1,33σ	11,9	13,3	13,3	22,5
20	73	79	69	83		Размах	34	36	36	54
21						6σ	53,7	60,0	60,0	101,3
22						Доля наблюдений в окрестности ±σ	68%	74%	74%	58%
23						Доля наблюдений в окрестности ±2σ	100%	95%	95%	100%
24										

Рис. 1. Оценки, полученные студентами при сдаче четырех тестов; мода зачеркнута, так как не имеет смысла

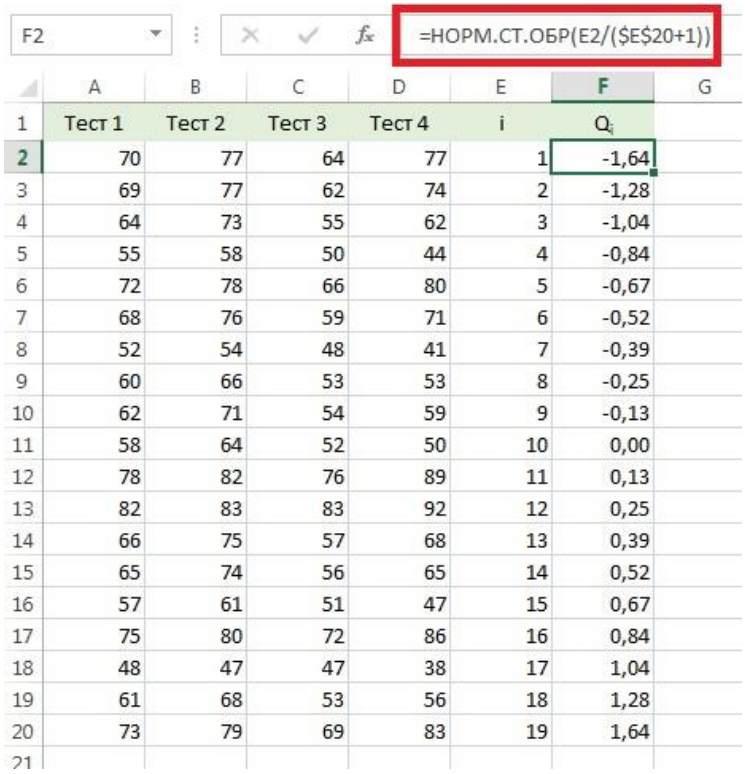
[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 368–375

Приблизительно нормальным является распределение оценок только по первому тесту: математическое ожидание равно медиане, доля наблюдений в пределах окрестности ±1σ от математического ожидания составляет 68% (в точности, как и для нормального распределения), асимметричность = 0.

Построение графика нормального распределения

Второй подход к проверке гипотезы о нормальном распределении использует график. Напомню, что для оценки смещения распределения были введены квартили. Кроме квартилей, для оценки нормальности распределения можно вычислять децили (разбивающие диапазон изменения данных на десятые доли), процентили (разбивающие диапазон изменения данных на сотые доли) и квантили (от слова *квант*), разбивающие всю совокупность данных на *n* диапазонов.

Для вычисления квантилей используется следующее правило (рис. 2): *i*-ый квантиль стандартизованного нормального распределения Q_i представляет собой стандартизованную нормально распределенную величину *Z*, которой соответствует площадь фигуры, лежащей под кривой плотности вероятностей, равная *i/(n+1)*.



	A	B	C	D	E	F	G
1	Тест 1	Тест 2	Тест 3	Тест 4	i	Q_i	
2	70	77	64	77	1	-1,64	
3	69	77	62	74	2	-1,28	
4	64	73	55	62	3	-1,04	
5	55	58	50	44	4	-0,84	
6	72	78	66	80	5	-0,67	
7	68	76	59	71	6	-0,52	
8	52	54	48	41	7	-0,39	
9	60	66	53	53	8	-0,25	
10	62	71	54	59	9	-0,13	
11	58	64	52	50	10	0,00	
12	78	82	76	89	11	0,13	
13	82	83	83	92	12	0,25	
14	66	75	57	68	13	0,39	
15	65	74	56	65	14	0,52	
16	57	61	51	47	15	0,67	
17	75	80	72	86	16	0,84	
18	48	47	47	38	17	1,04	
19	61	68	53	56	18	1,28	
20	73	79	69	83	19	1,64	
21							

Рис. 2. Расчет квантилей в Excel

График нормального распределения строится в Excel на основе точечного графика, на вертикальной оси которого отложены значения наблюдаемых данных, а на горизонтальной оси — соответствующие квантили стандартизованного нормального распределения (рис. 3). В отличие от описательных статистик, для построения графиков данные должны быть упорядочены по возрастанию. Если точки, соответствующие наблюдаемым данным, образуют прямую, проведенную из левого нижнего угла в правый верхний угол, значит, данные распределены приближенно нормально. С другой стороны, если эти точки отклоняются от прямой линии, распределение данных отличается от нормального.

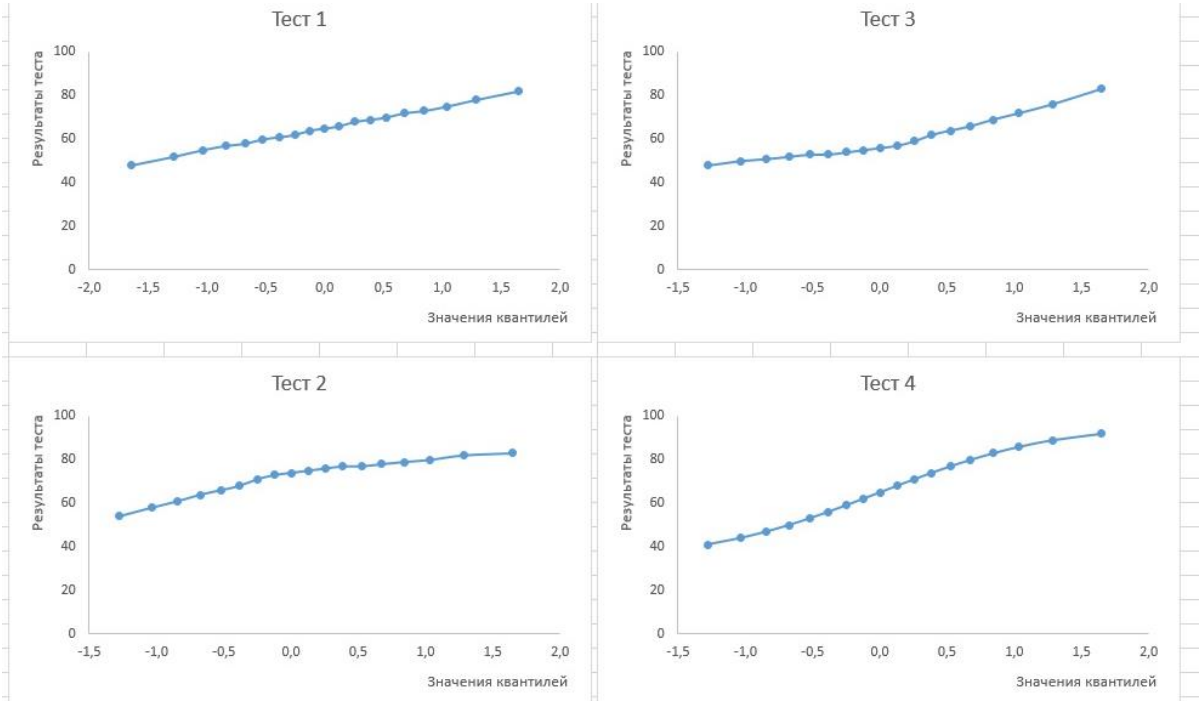


Рис. 3. Графики распределений для четырех тестов

График «Тест 1» свидетельствует, что наблюдаемые точки лежат очень близко к прямой линии, поэтому можно считать, что оценки, полученные студентами при сдаче первого теста, распределены практически нормально. Обратите внимание на полигон (кривую плотности распределения) и блочную диаграмму, изображенные на рис. 4, панель А.

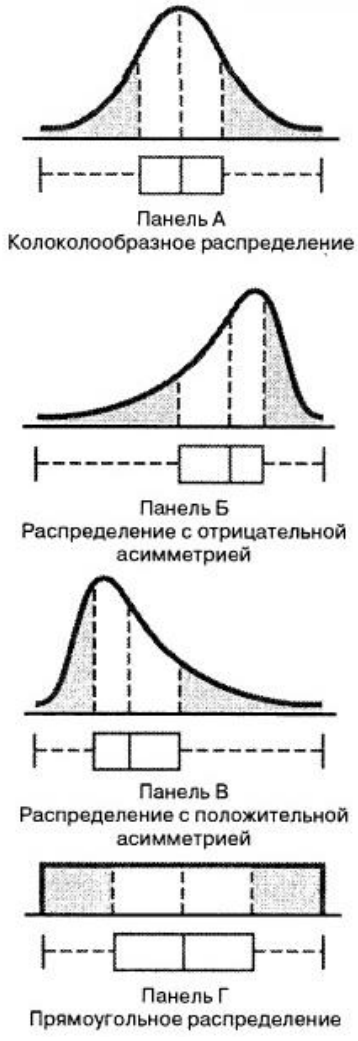


Рис. 4. Четыре распределения, исследованные с помощью блочных диаграмм

«Тест 2» (рис. 3): точки значительно отклоняются от прямой линии. Значения случайной переменной сначала возрастают довольно резко, а затем их рост становится умеренным. Этот рисунок соответствует распределению с отрицательной асимметрией, о чем свидетельствует более длинный левый хвост распределения. Обратите внимание на соответствующие полигон и блочную диаграмму, изображенные на рис. 4, панель Б. «Тест 3»: наблюдается противоположная картина. Значения случайной переменной сначала возрастают довольно медленно, а затем их рост становится более заметным. Этот рисунок соответствует распределению с положительной асимметрией, о чем свидетельствует более длинный правый хвост распределения. Обратите внимание на соответствующие полигон и блочную диаграмму, изображенные на рис. 4, панель В. «Тест 4»: изображен симметричный график, средняя часть которого почти линейна. Значения случайной переменной сначала довольно медленно возрастают, затем их рост прекращается, а в третьей части — ускоряется. Этот рисунок не совпадает ни с панелью Б, ни с панелью В. Это распределение не имеет хвостов. Следовательно, оно является равномерным (или прямоугольным). Обратите внимание на соответствующие полигон и блочную диаграмму, изображенные на рис. 4, панель Г.

Предыдущая заметка Нормальное распределение

Следующая заметка

К оглавлению Статистика для менеджеров с использованием Microsoft Excel