УДК 519.24

Расширение области применения критерия Ирвина при обнаружении аномальных измерений

В. В. Заляжных

Предложен способ оценки количества значений вариационного ряда, которые целесообразно проверять на аномальные наблюдения. Методами статистического моделирования получены таблицы процентных точек для критерия Ирвина, используемого при отбраковке аномальных наблюдений. Замена процентных точек критерия Ирвина при выборочном среднеквадратическом отклонении процентными точками при генеральном среднеквадратическом отклонении приемлема только при сравнительно больших объёмах выборок. Найдены уравнения, с приемлемой точностью аппроксимирующие процентные точки критерия Ирвина.

Ключевые слова: аномальные наблюдения, критерий Ирвина, процентные точки, статистическое моделирование.

1. Введение

В процессе получения данных, например, при измерениях, автоматической обработке и передаче сигналов и др., в выборке иногда получаются аномальные значения (грубые ошибки, промахи, выбросы), обусловленные изменением условий, сбоями средств измерений, ошибками оператора. Аномальными называют значения, резко отличающиеся по величине и статистическим свойствам от основной группы значений [1]. Доля аномальных значений может достигать 10 % и более [2–5]. Наличие их может приводить к существенным искажениям результатов обработки данных. Для отбраковки аномальных значений можно использовать различные статистические критерии. Большинство таких критериев основано на предположении, что выборка получена из нормально распределённой случайной величины. К таким критериям относится, в частности, критерий Ирвина. При этом нулевая гипотеза состоит в том, что все значения выборки принадлежат к одному и тому же нормальному распределению. Конкурирующая гипотеза состоит в том, что сомнительные значения принадлежат к некоторому другому распределению.

2. Применение критерия Ирвина

Критерий Ирвина предложен в 1925 г. в работе [6] для отбраковки аномальных значений из выборки, полученной из нормально распределённой случайной величины. В настоящее время критерий применяется при статистической обработке результатов измерений и испытаний [7–10], в частности, при предварительном анализе временных рядов, используемых при решении социально-экономических, технических и естественно-научных задач [11–17]. Применение критерия Ирвина рекомендовано также в ГОСТ 10518-88 «Системы электрической изоляции. Общие требования к методам ускоренных испытаний на нагревостойкость» и ГОСТ Р 51372-99 «Методы ускоренных испытаний на долговечность и сохраняемость при воздействии агрессивных и других специальных сред для технических изделий, материалов и систем материалов».

Порядок применения критерия Ирвина, исходя из [6], следующий. Значения выборки располагают в вариационный ряд: $x_1, x_2, ,,, x_i, \dots x_n$. Находят значение статистики Ирвина $\lambda_{расч}$ для k-го значения вариационного ряда по формуле

$$\lambda_{расч} = \frac{|x_k - x_{k+1}|}{\sigma}, \tag{1}$$

где k – номер значения x_i в вариационном ряду, считая с одного из концов ряда, σ – генеральное среднеквадратическое отклонение.

Поскольку σ обычно не известно с достаточной точностью, в [6] рекомендуется использовать выборочное среднеквадратическое отклонение s как лучшее приближение к нему. Эта рекомендация обычно используется на практике. Значение s рассчитывают по формуле

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

где $\bar{x}$ – среднее арифметическое значение выборки, n – объём выборки.

Проверку начинают с первого (с одного из концов ряда) значения ряда, т.е. при $k = 1$, и продолжают, переходя к следующим значениям ряда, при $k = 2$, затем $k = 3$ и т.д. Если для k-го значения ряда

$$\lambda_{расч} > \lambda_{k,\alpha},$$

где $\lambda_{k,\alpha}$ – табличное значение критерия Ирвина для k-го элемента ряда при выбранном уровне значимости α, то все значения ряда от первого до k-го считают аномальными и отбрасывают (бракуют). После этого следующие значения ряда проверяют в том же порядке, принимая для ближайшего последующего значения $k = 1$ и заново пересчитывая s. Такой подход выглядит обоснованным: если исключены промахи, оставшуюся выборку логично рассматривать как случайную и репрезентативную.

Проверять все значения ряда, очевидно, не следует, т.к. это может привести к ошибочной отбраковке большого количества значений. Целесообразно проверять некоторое предельное количество значений $k_{пр}$. Затем переходить к другому концу ряда и проверять также $k_{пр}$ значений. Далее чередовать проверку с разных концов ряда до тех пор, пока на обоих концах не будут отсутствовать аномальные значения. При этом общее количество проверенных значений на каждом конце ряда не должно превышать $k_{пр}$.

3. Проблемы применения критерия Ирвина и задачи исследования

Описанный алгоритм применения критерия Ирвина позволяет оценивать на аномальность большое количество элементов выборки. Однако табличные значения (процентные точки) критерия Ирвина приведены в [6] только для первого и второго значений вариационного ряда, притом для «некруглых» уровней значимости. По данным [6] рассчитаны и приведены в [19] табличные значения критерия $\lambda_{1,\alpha}$ (для первого элемента вариационного ряда) для «круглых» уровней значимости. Они воспроизведены во многих информационных источниках. Кроме того, в [20] приведены табличные значения критерия Ирвина для первого и второго элементов ряда. Но рассчитаны эти значения в [19, 20] для случая известного генерального среднеквадратического отклонения σ, а применяются обычно при использовании выборочного среднеквадратического отклонения s. В данной работе рассмотрена приемлемость такой замены.

Табличные значения критерия Ирвина для последующих значений вариационного ряда – третьего, четвёртого и т.д. – в информационных источниках отсутствуют. Но, как отмечено в [21] (по отношению к критерию Граббса), нет гарантии, что в выборке не большее число аномальных измерений, чем мы исследуем на выбросы, и это может отрицательно сказаться на результатах анализа данных. В данной работе предложен способ оценки количества значений выборки $k_{пр}$, которые целесообразно проверять на аномальность, и рассчитаны соответствующие табличные значения критерия Ирвина при выборочном среднеквадратическом отклонении для некоторых объёмов выборок в пределах от 3 до 1000.

Рассчитать табличные значения критерия Ирвина при всех возможных объёмах выборки (например, вплоть до 1000) затруднительно, и это вызывает определённые трудности при использовании критерия, особенно при автоматизированной обработке. В данной работе рассмотрена возможность аппроксимации табличных значений уравнениями, позволяющими рассчитывать процентные точки критерия при любых объёмах выборки вплоть до 1000.

4. Методика исследования

Определение процентных точек проводили методом статистического моделирования [22] с использованием табличного процессора MS Excel, реализуя при этом макрос, содержащий циклическую структуру. Процентные точки рассчитывали для уровней значимости 0.005, 0.01 и 0.05. Моделировали выборки из нормального стандартного распределения (использование нормальных распределений с иными параметрами даёт те же результаты) в количестве 1 млн при каждом конкретном объёме выборки.

По каждой выборке рассчитывали статистики (1) при k от 1 до $k_{пр}$, начиная с наибольшего значения вариационного ряда и используя при этом s вместо σ (расчёты от наименьшего значения ряда дают те же процентные точки). В совокупности статистик для $k = 1$ не учитывали наибольшие значения в доле, равной уровню значимости (соответствующие выборки не учитывались и в последующих расчётах при более высоких значениях k, поскольку на практике в таких выборках бракуются и отбрасываются наибольшие значения, что уменьшает объём выборки). Наибольшее из остальных значений статистики (1) принимали равным соответствующей процентной точке. Для последующих k процентные точки находили аналогичным образом.

Те смоделированные выборки объёмом n, из которых на некотором шаге k были отброшены аномальные значения (с вероятностью, равной уровню значимости), уменьшаются в объёме до $n-k$. Но их уже нельзя считать случайными и репрезентативными, соответствующими стандартному нормальному распределению. Поэтому при расчёте процентных точек для выборок объёмом $n-k$ их не учитывали. На практике такие выборки проверяются и далее на наличие аномальных значений. Возможные при этом результаты рассмотрены далее.

Таким образом, на каждом шаге количество используемых выборок N уменьшается в соответствии с соотношением

$$N = 10^6 \cdot (1-\alpha)^{k-1}.$$

Наименьшее количество выборок, используемых при расчёте процентных точек, составляет 487675 при $\alpha = 0.05$ и $k = 15$.

Для каждой найденной на уровне α процентной точки 99.73-процентный доверительный интервал для истинного уровня определяется выражением, приведённом в [23]:

$$\alpha \pm 3\sqrt{\alpha(1-\alpha)/N}.$$

Некоторые доверительные интервалы:

для $k = 1$ 0.005±0.0002. 0.01±0.0003. 0.05±0.0007;
для $k = 5$ 0.005±0.0002. 0.01±0.0003. 0.05±0.0007;
для $k = 10$ 0.005±0.0002. 0.01±0.0003. 0.05±0.0008;
для $k = 15$ 0.005±0.0002. 0.01±0.0003. 0.05±0.0009.

5. Оценка количества проверяемых значений

Если в выборке нет аномальных значений, то с вероятностью, равной уровню значимости, аномальность может быть ошибочно определена даже вблизи середины вариационного ряда, например, при автоматизированной обработке данных. Это совершенно неприемлемо, т.к. приведёт к ошибочному исключению из выборки большого количества значений. Поэтому количество значений, проверяемых на аномальность, следует ограничить в зависимости от объёма выборки. Для этого можно задаться вероятностью промаха p в каждом отдельном наблюдении. Такую вероятность можно принять, например, исходя из аналогичных серий измерений. Тогда вероятность h того, что промах встретится ровно m раз в выборке объёмом n, определяется формулой Бернулли:

$$h = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}. \quad (2)$$

Примем p = 0.005, что выглядит в общем случае вполне приемлемо. Тогда расчёты по формуле (2) показывают, что при конкретном n в пределах от 3 до 1000 вероятность h в целом снижается с ростом m и при некотором m становится очень мала. Если задаться некоторым предельным, достаточно маленьким, значением h, равным $h_{пр}$, то полученное значение m будет соответствовать значению $k_{пр}$, далее которого нецелесообразно продолжать проверку. Учитывая, что аномальные значения могут оказаться только на одном конце вариационного ряда, необходимо знать табличные значения критерия Ирвина λ_k при k от 1 до $k_{пр}$ для данного объёма выборки и уровня значимости.

При вероятности появления промаха в отдельном наблюдении p = 0.005 и заданном предельном значении $h_{пр}$ с использованием формулы (2) найдены предельные значения $k_{пр}$, приведённые в табл. 1.

Таблица 1. Предельные значения k_n количества проверяемых значений при p = 0.005

n	$h_{пр}$ = 0.0001	$h_{пр}$ = 0.0002	n	$h_{пр}$ = 0.0001	$h_{пр}$ = 0.0002
	$k_{пр}$			$k_{пр}$	
3	1	1	100	5	4
4	2	1	200	6	6
5	2	2	300	8	7
15	2	2	400	9	8
20	3	2	500	10	10
30	3	3	600	11	11
40	5	3	700	12	12
50	4	3	800	13	12
60	4	4	900	14	13
90	4	4	1000	15	14

6. Табличные значения критерия Ирвина

Наибольший практический интерес представляют табличные значения критерия Ирвина при расчёте среднеквадратического отклонения по выборке. С учётом данных табл. 1 рассчитаны такие табличные значения, приведенные в табл. 2 для уровня значимости 0.005, в табл. 3 – для уровня значимости 0.01, в табл. 4 – для уровня значимости 0.05.

Таблица 2. Табличные значения критерия Ирвина при расчёте среднеквадратического отклонения по выборке для уровня значимости 0.005

n	Номер промаха в вариационном ряду, k					n	Номер промаха в вариационном ряду, k				
	1	2	3	4	5		1	2	3	4	5
3	1.618					80	1.675	0.997	0.735	0.593	0.503
4	1.916	1.644				90	1.656	0.978	0.721	0.580	0.490
5	2.004	1.587				100	1.647	0.973	0.709	0.572	0.482
6	2.04	1.538	1.428			200	1.540	0.895	0.645	0.512	0.430

7	2.049	1.495	1.333			300	1.493	0.862	0.619	0.487	0.405
8	2.050	1.457	1.262			500	1.443	0.816	0.583	0.458	0.380
9	2.041	1.425	1.209			1000	1.375	0.772	0.548	0.428	0.353
10	2.028	1.395	1.164	1.063							
11	2.019	1.370	1.131	1.017		*n*	Номер промаха в вариационном ряду, *k*				
12	2.006	1.343	1.098	0.979			6	7	8	9	10
13	1.990	1.330	1.069	0.947		100	0.421	0.374	0.340	0.312	0.290
14	1.980	1.309	1.050	0.916		200	0.370	0.330	0.297	0.271	0.249
15	1.973	1.293	1.034	0.896		300	0.352	0.308	0.278	0.252	0.231
20	1.919	1.225	0.958	0.811		500	0.327	0.288	0.257	0.232	0.214
25	1.881	1.183	0.909	0.763		1000	0.301	0.265	0.236	0.213	0.195
30	1.841	1.149	0.873	0.730		*n*	Номер промаха в вариационном ряду, *k*				
35	1.818	1.120	0.847	0.702			11	12	13	14	15
40	1.792	1.099	0.824	0.679		200	0.202	0.189	0.177	0.168	0.160
50	1.749	1.061	0.793	0.650	0.556	300	0.187	0.174	0.163	0.154	0.146
60	1.719	1.036	0.768	0.623	0.535	500	0.172	0.160	0.150	0.141	0.133
70	1.699	1.014	0.750	0.603	0.516	1000	0.157	0.145	0.136	0.127	0.121

Таблица 3. Табличные значения критерия Ирвина при расчёте среднеквадратического отклонения по выборке для уровня значимости 0.01

n	Номер промаха в вариационном ряду, *k*					*n*	Номер промаха в вариационном ряду, *k*				
	1	2	3	4	5		1	2	3	4	5
3	1.722					80	1.498	0.885	0.649	0.523	0.443
4	1.879	1.604				90	1.481	0.869	0.636	0.512	0.433
5	1.932	1.519				100	1.471	0.861	0.628	0.503	0.425
6	1.944	1.454	1.349			200	1.374	0.792	0.569	0.451	0.378
7	1.932	1.399	1.244			300	1.329	0.759	0.544	0.429	0.355
8	1.917	1.354	1.169			500	1.281	0.720	0.513	0.403	0.333
9	1.900	1.317	1.110			1000	1.219	0.679	0.480	0.375	0.309
10	1.881	1.282	1.066	0.975							
11	1.865	1.257	1.032	0.926		*n*	Номер промаха в вариационном ряду, *k*				
12	1.848	1.228	1.000	0.892			6	7	8	9	10
13	1.830	1.212	0.972	0.859		100	0.371	0.331	0.300	0.274	0.255
14	1.817	1.190	0.953	0.831		200	0.326	0.290	0.260	0.237	0.219
15	1.804	1.174	0.933	0.809		300	0.308	0.271	0.243	0.221	0.203
20	1.743	1.108	0.862	0.728		500	0.287	0.252	0.224	0.204	0.187
25	1.700	1.061	0.814	0.681		1000	0.264	0.231	0.206	0.186	0.170
30	1.661	1.028	0.781	0.649		*n*	Номер промаха в вариационном ряду, *k*				
35	1.632	1.001	0.755	0.625			11	12	13	14	15
40	1.609	0.979	0.735	0.604		200	0.203	0.190	0.178	0.169	0.161
50	1.571	0.946	0.704	0.574	0.491	300	0.187	0.176	0.165	0.154	0.146
60	1.542	0.922	0.681	0.552	0.472	500	0.173	0.160	0.151	0.142	0.134
70	1.520	0.900	0.665	0.535	0.456	1000	0.157	0.146	0.136	0.128	0.120

Таблица 4. Табличные значения критерия Ирвина при расчёте среднеквадратического отклонения по выборке для уровня значимости 0.05

n	Номер промаха в вариационном ряду, *k*					*n*	Номер промаха в вариационном ряду, *k*				
	1	2	3	4	5		1	2	3	4	5
3	1.677					80	1.049	0.609	0.444	0.356	0.302
4	1.699	1.433				90	1.032	0.597	0.434	0.348	0.294
5	1.654	1.277				100	1.021	0.591	0.428	0.341	0.288
6	1.599	1.176	1.097			200	0.948	0.540	0.385	0.304	0.254
7	1.550	1.105	0.982			300	0.915	0.513	0.367	0.287	0.239
8	1.506	1.047	0.905			500	0.875	0.486	0.345	0.270	0.223
9	1.473	1.004	0.848			1000	0.831	0.457	0.320	0.250	0.206
10	1.442	0.969	0.803	0.738							
11	1.414	0.938	0.769	0.694		*n*	Номер промаха в вариационном ряду, *k*				
12	1.390	0.913	0.739	0.662			6	7	8	9	10
13	1.371	0.892	0.716	0.633		100	0.251	0.224	0.203	0.186	0.173
14	1.352	0.872	0.696	0.608		200	0.219	0.193	0.175	0.159	0.146
15	1.336	0.856	0.679	0.590		300	0.206	0.181	0.162	0.147	0.136
20	1.269	0.795	0.614	0.521		500	0.191	0.168	0.150	0.135	0.124
25	1.225	0.753	0.575	0.480		1000	0.175	0.154	0.136	0.124	0.113
30	1.192	0.725	0.548	0.454		*n*	Номер промаха в вариационном ряду, *k*				
35	1.164	0.701	0.526	0.433			11	12	13	14	15
40	1.143	0.684	0.511	0.418		200	0.135	0.127	0.120	0.113	0.108
50	1.111	0.658	0.486	0.395	0.338	300	0.125	0.117	0.110	0.104	0.0979
60	1.084	0.638	0.469	0.378	0.323	500	0.115	0.107	0.100	0.0941	0.0890
70	1.063	0.623	0.455	0.367	0.311	1000	0.104	0.0966	0.090	0.0848	0.0798

Рассчитаны табличные значения критерия Ирвина для некоторых n и k при известном генеральном среднеквадратическом отклонении σ (см. табл. 5). Здесь же приведены фактические уровни значимости, имеющие место, когда эти табличные значения используются, как это рекомендовано в [6], при замене в (1) σ на s.

Таблица 5. Табличные значения критерия Ирвина при известном среднеквадратическом отклонении

n	*α*	Табличные значения					Фактические уровни значимости				
		Номер промаха в вариационном ряду, *k*									
		1	2	3	4	5	1	2	3	4	5
4	0.005	2.854	2.394				0	0.00017			
	0.01	2.597	2.172				0	0.0002			
	0.05	1.918	1.597				0.0047	0.011			
10	0.005	2.264	1.537	1.279	1.166		0.0013	0.0018	0.0019	0.0020	
	0.01	2.047	1.380	1.144	1.039		0.0045	0.0055	0.0058	0.006	
	0.05	1.465	0.975	0.805	0.733		0.0466	0.0487	0.0497	0.0513	
30	0.005	1.900	1.176	0.898	0.744		0.004	0.0042	0.0042	0.0044	
	0.01	1.703	1.044	0.795	0.659		0.0086	0.0091	0.009	0.0092	
	0.05	1.197	0.725	0.547	0.451		0.0491	0.05	0.0503	0.0511	
70	0.005	1.716	1.022	0.755	0.610	0.520	0.0047	0.0048	0.0048	0.0047	0.0048
	0.01	1.531	0.906	0.667	0.539	0.457	0.0095	0.0097	0.0098	0.0096	0.0098

	0.05	1.067	0.621	0.454	0.365	0.310	0.0493	0.0507	0.0506	0.0505	0.0507
100	0.005	1.652	0.979	0.713	0.575	0.483	0.0049	0.048	0.0048	0.0049	0.0049
	0.01	1.475	0.865	0.630	0.504	0.425	0.0099	0.0098	0.0098	0.0098	0.0099
	0.05	1.022	0.590	0.426	0.341	0.286	0.0499	0.0502	0.0506	0.0505	0.0509

Табличные значения для k = 1 и k = 2, приведённые в табл. 5, очень близки или равны приведённым в [19, 20].

Сравнение таблиц 2, 3 и 4 с таблицей 5 показывает, что при относительно небольших n табличные значения для генерального среднеквадратического отклонения существенно больше, чем для выборочного. Соответственно, фактические уровни значимости значительно меньше принятых α. С увеличением n и k различия уменьшаются, проявляется тенденция к превышению фактических уровней значимости по сравнению с заданными (для α = 0.05) при увеличении k.

Определённый интерес представляет информация о том, начиная с каких объёмов выборки табличные значения при генеральном среднеквадратическом отклонении, приведённые в [19, 20], могут с приемлемой точностью применяться при расчёте статистик (1) по выборочному среднеквадратическому отклонению. Если принять допустимое отклонение фактического уровня значимости от принятого равным 10 %, то, как показывает расчёт, для первого значения вариационного ряда (k = 1) допустима замена выборочного среднеквадратического отклонения генеральным при объёмах выборки не менее:

n = 55 при α = 0.005 (фактически α = 0.0045);

n = 45 при α = 0.01 (фактически α = 0.009);

n = 9 при α = 0.05 (фактически α = 0.046).

7. Аппроксимация табличных значений

Для табличных значений критерия Ирвина при выборочном среднеквадратическом отклонении, приведённых в табл. 2 – 4, при объёмах выборок от 15 до 1000 подобрали аппроксимирующие уравнения вида

$$A\cdot(k-5/n)^{B}. \qquad (3)$$

Здесь k – номер значения в вариационном ряду с одного из концов ряда, n – объём выборки. Значения A и B определяются по выражениям, приведённым в табл. 6. Там же приведены максимальные Δ_{max} и средние $\Delta_{ср}$ ошибки аппроксимации. Ошибки аппроксимации рассчитывали как модули разностей между табличными значениями критерия Ирвина и рассчитанными по уравнениям вида (3). Приведены также δ_α – относительные отклонения уровней значимости от действительных при значениях критерия Ирвина, соответствующих Δ_{max}.

Таблица 6. Выражения для A и B в уравнениях вида (2), ошибки аппроксимации, отклонения уровней значимости

α = 0.005	A	$-114.686n^{0.2} + 615.0104n^{0.15} - 1234.813n^{0.1} + 1098.7951n^{0.05} - 363.701$
	B	$137.269n^{0.2} - 728.202n^{0.15} + 1450.2666n^{0.1} - 1285.8577n^{0.05} + 427.693$
	Δ_{max} = 0.007, $\Delta_{ср}$ = 0.0029, δ_α = 4.9 %	
α = 0.01	A	$-405.1713n^{0.25} + 2520.6255n^{0.2} - 6237.5919n^{0.15} + 7670.2996n^{0.1} - 4684.809n^{0.05} + 1138.003$
	B	$106.29403n^{0.2} - 569.75407n^{0.15} + 1146.93404n^{0.1} - 1028.2898n^{0.05} + 345.8343$
	Δ_{max} = 0.004, $\Delta_{ср}$ = 0.0013, δ_α = 3.5 %	
α = 0.05	A	$-4.041n^{0.5} + 32.5148n^{0.4} - 103.5032n^{0.3} + 162.9495n^{0.2} - 127.32n^{0.1} + 40.7683$
	B	$-0.30595n^{0.5} + 6.7127n^{0.4} - 38.1211n^{0.3} + 93.2983n^{0.2} - 106.1212n^{0.1} + 45.5395$
	Δ_{max} = 0.004, $\Delta_{ср}$ = 0.0011, δ_α = 2.7 %	

Уравнения вида (3) позволяют рассчитывать табличные значения критерия Ирвина при любых n в пределах от 15 до 1000 и при k, не превышающих тех, которые имеются для нижнего значения n в интервале, в котором находится n. Например, при $n = 150$ расчётные формулы с достаточной надёжностью применимы для k не более 10, поскольку в этом случае нижнее значение интервала n в табл. 2–4 равно 100 и для него имеются табличные значения для k до 10.

Для некоторых значений k и промежуточных значений n, не представленных в табл. 2–4, были найдены процентные точки методом статистического моделирования и сопоставлены с расчётными значениями, полученными по (3). При этом ошибки аппроксимации не превышали значений $\Delta_{\max}$, что подтверждает применимость уравнений вида (3) для любых объёмов выборок от 15 до 1000.

8. Выборки с ошибочно отбракованными аномальными значениями

Смоделированные выборки, из которых отброшены аномальные значения, как было указано выше, в дальнейших расчётах процентных точек не учитывали. На практике такие усечённые выборки, из которых ошибочно отброшены аномальные значения с вероятностью, равной уровню значимости, исследуются на наличие аномальных значений и далее, но уже при меньших n. При этом используются те же процентные точки, что и для других выборок. Но вероятности ошибок первого рода для ошибочно усечённых выборок не будут равны принятым уровням значимости. Очевидно, наибольшие различия имеют место для выборок наименьших объёмов.

Рассмотрены выборки начального объёма при $n = 5$ после ошибочного удаления наибольшего значения при уровне значимости 0.05. Для таких выборок дальнейшая проверка на аномальность при $n = 4$ для наибольшего из оставшихся значений дала вероятность ошибки первого рода 0.026, а для наименьшего значения – 0.101. Повышение этой ошибки по сравнению с принятым уровнем значимости для наименьшего значения полезно – удаление из выборки наибольшего и наименьшего значений обычно не приводит к существенному изменению среднего арифметического значения.

9. Заключение

Найденные процентные точки критерия Ирвина при выборочном среднеквадратическом отклонении и аппроксимирующие их уравнения позволяют более широко и статистически более обоснованно применять критерий на практике, в частности, при автоматизированной обработке данных.

Литература

1. *Марчук В. И., Токарева С. В.* Способы обнаружения аномальных значений при анализе нестационарных случайных процессов: монография. Шахты: Изд. ЮРГУЭС, 2009. 209 с.
2. *Хогг Р. В.* Введение в помехоустойчивое оценивание / В кн.: Устойчивые статистические методы оценки данных. Пер.с англ. М.: Машиностроение, 1984. С. 86–105.
3. *Аджи У. С., Тернер Р. Х.* Применение методов помехоустойчивого оценивания в анализе данных о траекториях движения / В кн.: Устойчивые статистические методы оценки данных. Пер.с англ. М.: Машиностроение, 1984. С. 12–26.
4. *Патюков В. Г.* Основы частотно-временных измерений: монография. Красноярск: Сиб. федер. ун-т, 2014. 166 с.

5. *Серышева И. А.* Фильтрация выбросов в задачах статической и динамической обработки данных в эталонах времени и частоты // Вестник Иркутского государственного технического университета. 2018. Т. 22, № 10. С. 67–77.
6. *Irvin J. O.* On a criterion for the rejection of outlying observation // Biometrika. 1925. V. 17. P. 238–250.
7. *Гергет О. М., Константинова Л. И., Кочегуров В. А.* Математические методы прогнозирования здоровья детей раннего возраста // Успехи современного естествознания. 2013. № 5. С. 165–169.
8. *Попукайло В. С.* Исследование критериев грубых ошибок применительно к выборкам малого объема // Радіоелектронні і комп'ютерні системи. 2015. № 3. С. 39–44.
9. *Попукайло В. С.* Обнаружение аномальных измерений при обработке данных малого объема // Технология и конструирование в электронной аппаратуре. 2016. № 4–5. С. 42–46.
10. *Кобзарь А. И.* Прикладная математическая статистика. М.: ФИЗМАТЛИТ, 2006. 816 с.
11. *Горбун Е. С., Максимов Н. В., Монанков К. В., Низаметдинов Ш. У.* Модель и средства интерактивного анализа динамики и связей потоков публикаций научной информации // Научная визуализция. 2015. Т. 4, № 5. С. 12–25.
12. *Трофименко С. В., Трофименко С. В., Маршалов А. Я., Гриб Н. Н., Колодезников И. И.* Модификация метода Ирвина для выявления аномальных уровней временных рядов: методика и численные эксперименты // Современные проблемы науки и образования. 2014. № 5. С. 255.
13. *Пшеничнов Р. В.* Анализ факторов, влияющих на результаты банковской деятельности в сфере ипотечного кредитования // Жилищные стратегии. 2017. Т. 4, № 2. С. 107–126.
14. *Архипова А. В.* Применение критерия Ирвина для анализа динамики прямых иностранных инвестиций в Российской Федерации // Juvenis scientia. 2015. № 1. С. 43–45.
15. *Копелиович Д. И., Боровикова В. О.* Применение анализа временных рядов для прогнозирования распределения средств по государственным контрактам в здравоохранении. // Вестник Брянского государственного технического университета. 2013. № 2 (38). С. 106–115.
16. *Бунтова Е. В.* Модель прогнозирования показателя инфляции на 2016–2018 год // Международный журнал прикладных и фундаментальных исследований. 2016. № 6–2. С. 357–360.
17. *Бучацкая В. В.* Обработка аномальных значений уровней временного ряда как этап комплексной оценки информации в подсистеме прогнозирования для ситуационного центра // Вестник Адыгейского государственного университета. Сер.: Естественно-математические и технические науки. 2013. В. 3. С. 98–102.
18. *Горчаков А. А.* Математический аппарат для инвестора // Аудит и финансовый анализ. 1997. № 3. С. 161–215.
19. *Янко Я.* Математико-статистические таблицы / Пер. с чеш. М.: Госстатиздат, 1961. 243 с.
20. *Barnett V., Lewis T.* Outliers in Statistical Data. John Wiley: Wiley, 1978. 365 p.
21. *Лемешко Б. Ю., Лемешко С. Б.* Расширение области применения критериев типа Граббса, используемых при отбраковке аномальных измерений // Измерительная техника. 2005. № 6. С. 13–19.
22. *Лемешко Б. Ю., Лемешко С. Б., Постовалов С. Н., Чимитова Е. В.* Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход: монография. Новосибирск: Изд-во НГТУ, 2011. 888 с.
23. *Zhang J.* Powerful goodness-of-fit and multi-sample tests // PhD Thesis. York University, Toronto, 2001.

Статья поступила в редакцию 22.02.2020;
переработанный вариант – 27.03.2020.

Заляжных Владимир Васильевич
к.т.н., доцент кафедры стандартизации, метрологии и сертификации Северного (Арктического) федерального университета им. М. В. Ломоносова (163002, Архангельск, наб. Северной Двины, 17), e-mail: `zalvladimir@yandex.ru`.

Widening of the application area of Irwin's test used for detections of outlying observations

V.V. Zaliazhnykh

Estimation of the number of the variation series values which should be checked for outlying observations is proposed. The tables of percentage points of Irwin's test used in the rejection of outlying observations are obtained by statistical simulation methods. Replacing the percentage points of the Irwin's test for the sample standard deviation with the percentage points for the standard-deviation of the original population is acceptable only for relatively large sample volumes. Equations are found, approximating the percentage points of the Irwin criterion with acceptable accuracy.

Keywords: outlying observations, Irwin's test, percentage points, statistic simulation.