Известия Гомельского государственного университета
имени Ф. Скорины, №5(74), 2012

УДК 58.087+581.5

Первичный анализ эмпирических распределений

Ю.М. Жученко, П.П. Евтухов

Для оценки статистических характеристик измеряемых величин включены: анализ резко выделяющихся наблюдений (выпадений или артефактов) и выбраковка недостоверной информации; определение закона и оценка параметров распределений; определение числа измерений, необходимых для оценки параметров выбранной вероятностной модели с заданной точностью.
Ключевые слова: статистические характеристики измеряемых величин, резко выделяющиеся наблюдения, выбраковка недостоверной информации, параметры распределения выборки, параметры распределения, вероятностная модель.

The estimation of statistical characteristics of measured values includes: the analysis of sharply allocated supervision (losses or artifacts) and unreliable information removal; definition of the law and an estimation of parameters of distributions; definition of number of the measurements necessary for an estimation of parameters of chosen probabilistic model with set accuracy.
Keywords: statistical characteristics of measured values, sharply allocated supervision, unreliable information removal, parameters of sample distribution, parameters of distributions, probabilistic model.

Введение

Методы и модели, применяемые для решения задач взаимосвязи измеряемых параметров и прогноза их изменений, базируются на экспериментальных данных, причем в качестве параметров моделей, как правило, используют средние величины. В то же время эти показатели характеризуются значительной вариабельностью, что обусловлено пространственной неоднородностью объекта исследований, погрешностями методического характера и при обработке исходной информации. Это обстоятельство обусловливает существенную вариабельность параметров моделей, что резко снижает точность прогноза. Таким образом, реалистичность прогноза можно повысить, привлекая информацию о статистических характеристиках эмпирических распределений измеряемых параметров. Статистический анализ эмпирических распределений исходных данных имеет и самостоятельную ценность, поскольку применение статистических методов позволяет отбраковать недостоверную информацию, повысить точность оценок и эффективность экспериментальных исследований путем определения оптимального объема экспериментальных данных.

Оценка статистических характеристик измеряемых величин включает:

– анализ резко выделяющихся наблюдений (выпадений или артефактов) и выбраковку недостоверной информации;

– определение закона и оценку параметров распределений;

– определение числа измерений, необходимых для оценки параметров выбранной вероятностной модели с заданной точностью.

Анализ резко выделяющихся наблюдений, выбраковка недостоверной информации

Наблюдения, которые резко выпадают из общего ряда (артефакты), могут быть обусловлены как ошибками при отборе и анализе проб, несоблюдением регламента эксперимента, изменением его условий, так и незамеченными или неучтенными существенными для понимания исследуемого процесса факторами. Поэтому анализ выделяющихся наблюдений позволяет еще раз проверить условия их регистрации и тем самым выделить и устранить ошибку.

Анализ резко выделяющихся наблюдений включает два этапа: выявление «подозрительных» наблюдений и проверку с помощью статистических критериев их принадлежности к принятой для описания основной массы экспериментальных данных вероятностной модели. Простейшим критерием на выявление резко выделяющихся наблюдений является «правило четырех сигм», которое гласит, что наблюдение может быть отброшено, если оно лежит вне

области $\mu \pm 4\sigma$, причем среднее значение μ и стандартное отклонение σ рассчитываются без учета проверяемых на выброс экспериментальных данных. Этот критерий основан на том, что «интервал четырех сигм» ($\mu \pm 4\sigma$) включает в себя при нормальном распределении 99,99% объема генеральной совокупности.

При малых объемах выборки ($n < 25$) проверка гипотезы о том, относится ли резко выделяющееся наблюдение к той же генеральной совокупности, что и остальные данные, осуществляется с помощью статистики Диксона:

$$M = \frac{X_1 - X_3}{X_1 - X_{n-2}}, \quad (1)$$

где X_1 – подозреваемое на выброс значение; X_3 и X_{n-2} – третье значение соответственно от начала и конца в упорядоченной по величине выборке. Критические значения статистики (1) приведены в работе [1].

Если объем выборки превосходит 25, то проверка экспериментальных значений на выброс осуществляется с помощью статистики:

$$T = \frac{X_1 - \mu}{\sigma}. \quad (2)$$

Если величина статистики (2) превосходит критическое значение [1], то анализируемое наблюдение считается выбросом.

Если в аномальном поведении подозревается сразу несколько значений, то статистики (1) или (2) сначала применяют к максимально выделяющемуся из подозреваемых на выброс значений. Если оно признается выбросом, то его удаляют из выборки, а статистики (1) или (2) применяют к следующему по величине значению, пока не будет признано, что выбросов больше нет.

Необходимо подчеркнуть, что статистики (1), (2) предполагают приближенно нормальное распределение основной массы данных. Это крайне важное требование, поскольку прямое математическое моделирование показало [2], что статистические процедуры, предполагающие нормальное распределение, неожиданно быстро теряют свои оптимальные свойства при утяжелении «хвостов» распределения. Есть основания полагать, что предположение нормальности при анализе эмпирических распределений исходных данных выполняется. Например, в работе [3] на основе анализа собственных экспериментальных материалов и данных литературы показано, что для широкого круга антропогенных загрязнителей распределение значений коэффициентов накопления удовлетворяет либо нормальному, либо логарифмически нормальному закону распределения.

Определение закона и оценка параметров распределения

Для получения надежных статистических оценок исходных данных, которые отвечают трем основным требованиям, предъявляемым к статистическим оценкам [1], – несмещенности, состоятельности и эффективности – крайне важное значение имеет правильная идентификация закона распределения экспериментальных данных. Так, известно, что при логарифмически нормальном распределении экспериментальных данных среднее арифметическое (оценка максимального правдоподобия в случае нормального распределения) является несмещенной и состоятельной, но не наиболее эффективной оценкой математического ожидания [1]. Поэтому до определения статистических характеристик измеряемых параметров необходимо идентифицировать закон распределения.

Задача проверки гипотезы о законе распределения случайной величины формулируется следующим образом. Пусть в распоряжении исследователя есть фиксированная выборка объемом n:

$$\tilde{O} = \tilde{o}_1,\ \tilde{o}_2,\ \ldots\ \tilde{o}_n.$$

Тогда нулевая гипотеза H_0 состоит в предположении, что X описывается распределением $F(x)$, а конкурирующая гипотеза H_1 заключается в том, что $F(x)$ не является функцией распределения X либо функцией распределения X является $F_1(x)$. Для проверки гипотезы H_0 против альтернативы H_1 строят критерий, основанный на использовании заранее определенной

меры расстояния между анализируемой эмпирической функцией распределения $F_n(x)$ и гипотетической модельной $F(x)$.

Объем экспериментальных данных в задаче оценки измеренных величин часто бывает небольшим. Поэтому для идентификации закона распределения желательно иметь чувствительный непараметрический критерий согласия, характеризующийся быстрой сходимостью к предельным распределениям. Таким критерием является более мощный в области малых выборок по сравнению с популярными критериями χ^2 и Колмогорова – Смирнова критерий ω^2 Мизеса [1]. Статистикой критерия является величина среднеквадратичного отклонения эмпирической функции распределения от гипотетической:

$$\omega_n^2 = n \cdot \int \left(F(x) - F_n(x) \right)^2 \cdot dF(x). \tag{4}$$

Хотя статистика (4) с ростом объема выборки быстро приближается к предельному распределению, для маленьких значений n расхождение между нею и предельным распределением может быть существенным. В этом случае используется модификация статистики ω_n^2, которая для малых n значительно лучше согласуется с предельным распределением. Приведем модификации статистики ω_n^2 отдельно для правого:

$$\tilde{\omega}^2 = \left(\omega_n^2 - \frac{0{,}4}{n} + \frac{0{,}6}{n^2} \right) \cdot \left(1{,}0 + \frac{1{,}0}{n} \right) \tag{5}$$

и левого «хвостов» распределения:

$$\tilde{\omega}^2 = \left(\omega_n^2 - \frac{0{,}03}{n} \right) \cdot \left(1{,}0 + \frac{0{,}5}{n} \right). \tag{6}$$

Основным недостатком статистики (4) является то, что гипотетическая функция распределения $F(x)$ должна быть полностью определена, вплоть до значений ее параметров.

Как показано в работе [5], если оценке подлежат параметры сдвига и масштаба (что и имеет место в случае нормального и логарифмически нормального законов распределения), предельное распределение статистики ω_n^2 будет зависеть только от формы распределения $F(x)$, но не от его параметров. Независимость предельного распределения от значений параметров сдвига и масштаба дает возможность построить на основе статистики (4) критерий проверки нормального характера распределения. При этом параметры гипотетического распределения $F(x)$ оцениваются по известным формулам:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}; \quad \sigma^2 = \frac{\sum_{i=1}^{v} \left(x_i - \mu \right)^2}{n-1}. \tag{7}$$

В работе [2] приведены критические значения статистики ω_n^2 для условий, когда параметры гипотетического распределения оцениваются по соотношениям (7).

Известно, что для нормального распределения среднее значение μ совпадает с модой x_{mod}, а для логарифмически нормального распределения имеем:

$$\frac{x_{\text{mod}}}{\mu} = \sqrt{V^2 + 1}, \tag{8}$$

где V – коэффициент вариации.

Из соотношения (8) следует, что при небольших значениях коэффициентов вариации законы нормального и логарифмически нормального распределения в плане средней величины и стандартных отклонений практически эквивалентны. Например, при V = 0,25 величина отклонения (8) составляет $\frac{x_{\text{mod}}}{\mu} \approx 1{,}03$, то есть допущение закона нормального распределения, хотя правильнее было бы применить логарифмически нормальное, вводит погрешность ≈ 3% по средней величине и моде. С практической точки зрения отсюда можно сделать следующий вывод: при небольшой величине V бесполезно увеличивать число наблюдений с целью более точного определения закона распределения, поскольку погрешность при выравнивании среднего и моды будет мала. В этом случае для оценки величины исследуемого параметра может

быть использовано среднее значение *x*. В то же время при значительной величине *V* необходимо иметь небольшое число наблюдений, чтобы точно идентифицировать закон распределения.

Определение числа проб, необходимых для оценки параметров вероятностной модели с заданной точностью

Одной из наиболее важных задач планирования экспериментов по определению параметров генеральной совокупности является оценка числа испытаний. Грамотно спланированным опытом можно считать такой, когда ответ на поставленный вопрос получается с наименьшими затратами сил и времени, а это прежде всего означает, что число повторностей в исследованиях должно быть минимальным. Знание закона распределения позволяет не только дать удовлетворительную оценку величины исследуемого параметра, но и определить число наблюдений, необходимых для получения значений параметра с фиксированной относительной вероятностной погрешностью при заданной доверительной вероятности.

Для выборочного среднего (7) из нормально распределенной генеральной совокупности с параметрами μ и σ с вероятностью $1-\alpha$ можно утверждать, что интервал $\mu - \frac{t \cdot \sigma}{\sqrt{n}}; \quad \mu + \frac{t \cdot \sigma}{\sqrt{n}}$ накрывает истинное значение $\overline{\mu}$. Рассматривая половину доверительного интервала как допустимую вероятную погрешность определения среднего, относительную вероятную погрешность запишем в виде:

$$\delta = \frac{\Delta}{\mu} = \frac{t \cdot \sigma}{\mu \cdot \sqrt{n}} = \frac{t \cdot V}{\sqrt{n}},$$

где t – значение критерия Стьюдента.

Таким образом, число проб, которое необходимо иметь для определения средних значений с относительной вероятной погрешностью δ, можно определить по формуле:

$$n = \left(\frac{t \cdot V}{\delta} \right)^2 . \qquad (9)$$

Принято считать, что для предварительной оценки измеряемой характеристики следует выбирать величину относительной вероятной погрешности в пределах 25–30 %, а при детальном изучении объекта исследований – не более 15–20%.

Из соотношения (9) видно, что с повышением доверительной вероятности $1-\alpha$ и снижением относительной вероятной погрешности δ число необходимых проб увеличивается. Рассчитанное по формуле (9) число проб, необходимых для оценки значений измеряемых параметров средним арифметическим в случае нормального распределения в зависимости от величины коэффициента вариации и относительной вероятной погрешности при $\alpha = 0{,}05$, представлено в таблице 1. График зависимости $n(V)$ для разных значений δ представлен на рисунке 1.

Таблица 1 – Минимальное количество проб, необходимых для оценки исследуемого параметра средним арифметическим (нормальное распределение, α = 0,05)

Коэффициент вариации	Относительная вероятная погрешность				
	0,1	0,15	0,2	0,25	0,3
0,1	6	4	4	3	3
0,2	18	9	6	5	4
0,3	37	18	11	8	6
0,4	64	30	18	12	9
0,5	98	45	26	17	13
0,6	142	64	37	25	18
0,7	189	86	49	32	23
0,8	248	112	64	42	30

Коэффициент вариации	Относительная вероятная погрешность				
	0,1	0,15	0,2	0,25	0,3
0,9	313	141	80	52	37
1,0	386	173	98	64	45
1,2	556	247	141	91	64
1,4	756	337	191	123	86
1,6	986	440	248	160	112
1,8	1247	556	314	201	141
2,0	1537	686	387	248	173

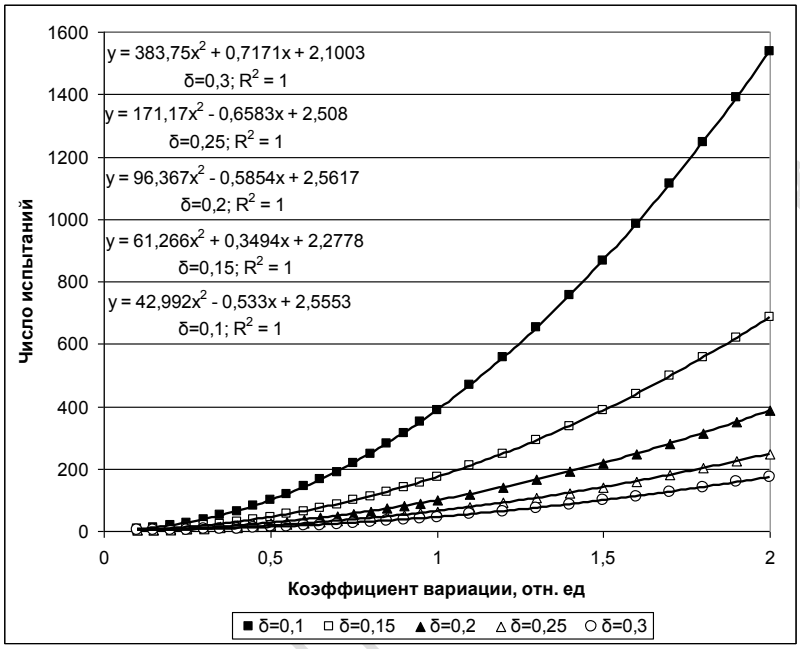

Рисунок 1 – Минимальное число проб (n), необходимое для оценки параметров вероятностной модели, основанной на нормальном распределении в зависимости от коэффициента вариации V и относительной вероятной погрешности δ

Видно, что при небольших значениях коэффициента вариации ($V<30\%$) для оценки среднего с относительной вероятной погрешностью ($\delta = 20\%$) и принятой доверительной вероятностью ($1–\alpha = 0,95$) необходимо отобрать и проанализировать менее 11 проб. С ростом V для получения оценки параметров с такой же относительной вероятной погрешностью число проб резко возрастает. В левом углу рисунка приведены аппроксимации полиномом второго порядка числа испытаний (y) при изменении коэффициента вариации (x) с фиксированным значением относительной вероятной погрешности δ.

Если распределение случайной величины отличается от нормального, то оценка центра распределения средним арифметическим, оставаясь состоятельной, перестает быть самой эффективной [4]. Более того, даже в случае симметричных распределений эффективность статистики μ как оценки параметра сдвига быстро падает с утяжелением «хвостов» распределения [2]. В случае асимметричного унимодального распределения, каким является логарифмически нормальное, для оценки центра распределении целесообразно использовать моду. При логарифмически нормальном распределении относительная вероятная погрешность связана с коэффициентом вариации и числом проб следующим образом [1]:

$$\delta^2 = \frac{\ln\left(V^2+1\right)\cdot\left(3n-1\right)}{n\cdot\left(n-1\right)}. \qquad (10)$$

Рассчитанное по формуле (10) число проб, необходимых для оценки значений параметров для логарифмически нормального распределения в зависимости от относительной вероят-

ной погрешности и коэффициента вариации, представлено в таблице 2. График зависимости *n(V)* для разных значений δ представлен на рисунке 2.

Таблица 2 – Минимальное количество проб, необходимых для оценки исследуемого параметра модой (логарифмически нормальное распределение, $\alpha = 0{,}05$)

Коэффициент вариации	Относительная вероятная погрешность				
	0,1	0,15	0,2	0,25	0,3
0,2	12	6	4	3	2
0,3	27	12	7	5	4
0,4	45	20	12	8	6
0,5	68	30	17	11	8
0,6	93	42	24	15	11
0,7	120	54	31	20	14
0,8	149	67	38	24	17
0,9	179	80	45	29	20
1	209	93	53	34	24
1,2	268	120	68	43	30
1,4	326	145	82	53	37
1,6	382	170	96	62	43
1.8	434	194	109	70	49
2	484	215	121	78	54

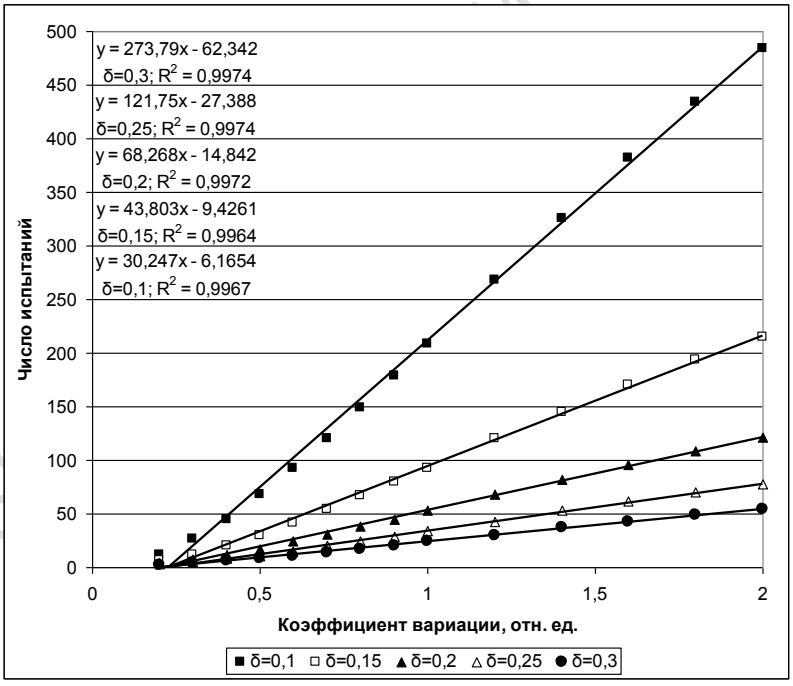

Рисунок 2 – Минимальное число проб (*n*), необходимое для оценки параметров вероятностной модели, основанной на логарифмически нормальном распределении в зависимости от коэффициента вариации *V* и относительной вероятной погрешности δ

В левом углу рисунка приведены аппроксимации линейной функцией числа испытаний (*y*) при изменении коэффициента вариации (*x*) с фиксированным значением относительной вероятной погрешности δ.

Сопоставляя объемы выборок, соответствующие одинаковым значениям δ и *V* (при $\alpha = 0{,}05$), видим, что в случае оценки модой часто требуется меньшее количество проб, чем при оценке средним арифметическим.

Таким образом, число точечных проб, достаточное для составления объединенной выборки, характеризующей с заданной точностью измеряемый параметр, определяется величиной коэффициента вариации и точностью.

Литература

1. Закс, Л. Статистическое оценивание / Л. Закс. – М. : Статистика, 1976. – 598 с.

2. Айвазян, С.А. Прикладная статистика: основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М., 1983. – 471 с.

3. Сатаева, Л.В. Применение статистических методов к решению некоторых вопросов загрязнения почв : автореф. канд. дис. / Л.В. Сатаева. – Обнинск, 1990. – 23 с.

4. Крамер, Г. Математические методы статистики / Г. Крамер. – М. : Мир, 1975. – 648 с.

5. Кендалл, М.Дж. Курс статистики. Статистические выводы и связи. Т. 2. / М.Дж. Кендалл, А. Стьюарт. – М. : Наука, Физматлит, 1973. – 466 с.

Гомельский государственный
университет им. Ф. Скорины

Поступило 15.05.12