УДК 519.281.1

К. А. РЕЗНИК

(Ленинград)

МЕТОД ИСКЛЮЧЕНИЯ РЕЗКО ВЫДЕЛЯЮЩИХСЯ НАБЛЮДЕНИЙ ДЛЯ ОДНОМОДАЛЬНЫХ РАСПРЕДЕЛЕНИЙ

В литературе по вопросам математической статистики описываются методы исключения резко выделяющихся наблюдений [1—3], основанные на гипотезе нормальности распределения генеральной совокупности наблюдений. Однако в ряде случаев распределение экспериментальных данных нельзя считать нормальным или же оно просто неизвестно. В связи с этим возникает вопрос, как исключить далеко отстоящие наблюдения при других формах распределений?

В данной работе предлагается метод исключения далеко отстоящих наблюдений для семейства распределений, описанного в [4—6], позволяющий одновременно с этим выбрать теоретическое распределение из этого семейства, наиболее близкое к экспериментальным данным. Критерием такой близости принят минимум среднего значения

$$\bar{\chi}^2 = \frac{1}{R}\sum_{i=1}^{L}\chi_i^2 = \frac{1}{R}\sum_{i=1}^{L}\frac{(n_i - \nu_i)^2}{\nu_i}, \qquad (1)$$

где $R=L-G$ — число степеней свободы; L — число интервалов группирования; G — число параметров теоретического распределения, оцениваемых по экспериментальным данным; n_i — число наблюдений в i-м интервале; ν_i — теоретическое значение частости i-го интервала.

Определение «наилучших» значений параметров распределений по минимуму χ^2 изложено в [7]. Переход к среднему значению $\bar{\chi}^2$, учитывающему число степеней свободы, позволяет не применять доверительные вероятности.

Рассмотрим некоторые свойства распределения экспериментальных данных. Пусть дано распределение $n+1$ результатов эксперимента, состоящее из компактной части, включающей n наблюдений, и одного далеко отстоящего наблюдения x_{n+1}. Пример такого распределения показан на рис. 1. Вычислим среднее арифметическое $\bar{x}_1$ и несмещенные оценки дисперсии M_{21} и эксцесса $\Э_1$ по всем экспериментальным данным, а также среднее арифметическое $\bar{x}_0$ и смещенные оценки $\widehat{M}_{20}$ $\widehat{\Э}_0$, не учитывая далеко отстоящего наблюдения.

Оценки дисперсии M_{21} и M_{20} будут связаны уравнением

$$M_{21} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_0)^2 + \frac{(x_{n+1} - \bar{x}_1)^2}{n} + (\bar{x}_1 - \bar{x}_0) = \widehat{M}_{20} + \frac{(x_{n+1} - \bar{x}_0)^2}{n+1}.$$

Обозначим $t_0 = (x_{n+1} - \bar{x}_0)/\sqrt{\widehat{M}_{20}}$. Тогда

$$\widehat{M}_{20} = M_{21}(1 - t_0^2/(n+1)). \qquad (2)$$

Для оценок эксцесса имеем

$$\Э_1 = \Э_0\frac{1}{\left(1+\frac{t_0^2}{n+1}\right)^2} + \frac{t_0^4 B_0}{(n+1)\left(1+\frac{t_0^2}{n+1}\right)^2},$$

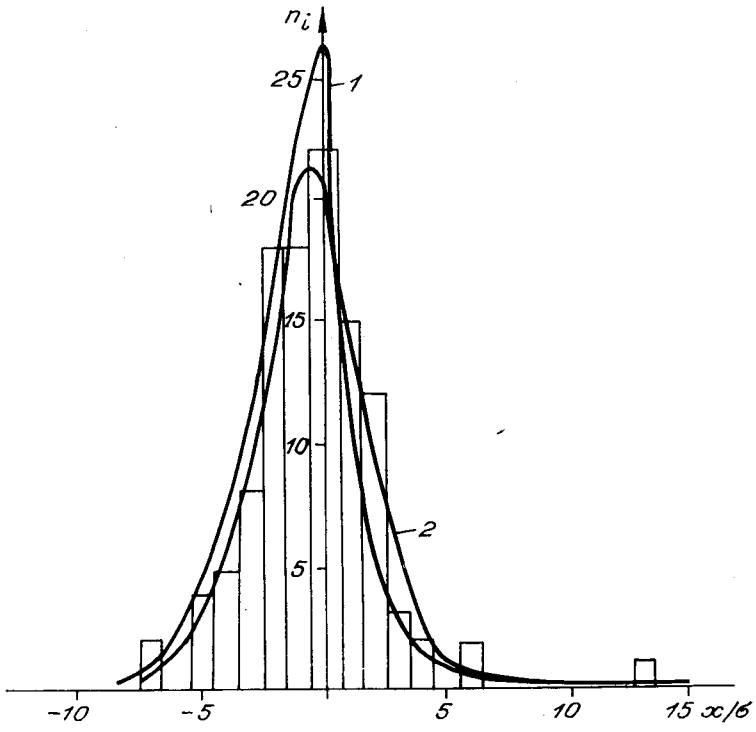

Рис. 1. Распределение экспериментальных данных.

где

$$B_0 = \frac{6\widehat{M}_{20}}{(n+1)\left(x_{n+1} - \bar{x}_0\right)^2} - \frac{4\widehat{M}_{30}}{x_{n+1} - \bar{x}_0} + \frac{n^3+1}{(n+1)^3};$$

при больших n $B_0 \approx 1$; $\widehat{M}_{30}$ — смещенная оценка третьего начального момента.

Отсюда

$$Э_0 = Э_1\left(1 + \left(t_0^2/n + 1\right)\right)^2 + \left(t_0^4 B_0/n + 1\right). \qquad (3)$$

Из формулы (2) видно, что при отбрасывании крайнего члена вариационного ряда, составленного из абсолютных отклонений от среднего арифметического, оценка дисперсии всегда уменьшается.

Из формулы (3) следует, что уменьшение оценки эксцесса при переходе от $Э_1$ к $Э_0$ происходит только в тех случаях, когда отбрасываемый $(n+1)$-й член ряда имеет достаточно большое нормированное отклонение, т. е. является далеко отстоящим. При большом числе данных ($n > 50$) смещенные оценки $\widehat{M}_{20}$ и $\widehat{Э}_0$ незначительно отличаются от несмещенных M_{20} и $Э_0$. Поэтому поправки в формулы (2) и (3) можно не вносить. Эти поправки, принципиально ничего не меняя, только усложнят выражения (2) и (3).

На рис. 2 показаны области различного изменения оценок эксцесса при отбрасывании одного наблюдения для различных значений Э и t_0 при n от 50 до 200.

Перейдем к теоретической модели распределения. Семейство одномодальных теоретических распределений [4, 5] можно представить уравнением

$$f(x, \sigma, K) = \frac{K\left[\Gamma\left(\frac{3}{K}\right)\right]^{1/2}}{2\sigma\left[\Gamma\left(\frac{1}{K}\right)\right]^{3/2}} \exp\left\{-\left[\frac{|x|}{\sigma}\left(\frac{\Gamma\left(\frac{3}{K}\right)}{\Gamma\left(\frac{1}{K}\right)}\right)^{1/2}\right]^K\right\}, \qquad (4)$$

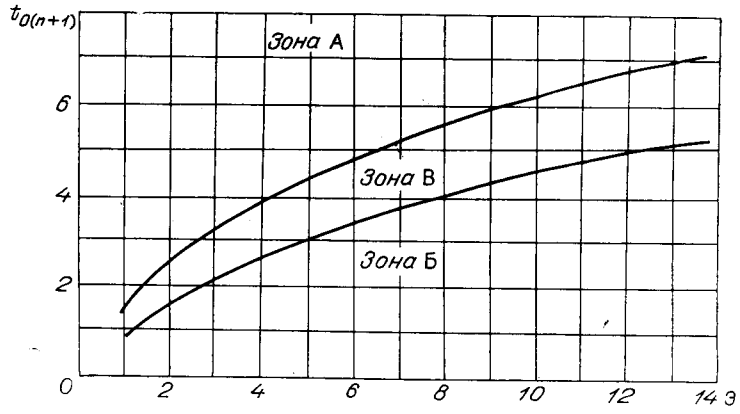

Рис. 2. Области изменения оценок эксцесса при отбрасывании одного наблюдения:

А — отбрасывание наблюдения уменьшает оценку эксцесса; Б — отбрасывание наблюдения увеличивает оценку эксцесса; В — зона неопределенности.

где σ — среднеквадратическое отклонение; K — постоянная. Подстановка в (4) $K=1$ дает выражение для плотности двойного экспоненциального распределения. При $K=2$ получаем выражение плотности нормального распределения. При $K \to \infty$ функция $f(x, \sigma, K)$ стремится к плотности равномерного распределения.

Выражение (4) позволяет классифицировать также промежуточные формы распределений. Кривые плотности таких распределений показаны на рис. 3. Значение K зависит от эксцесса

$$Э = \frac{\Gamma(1/K)\,\Gamma(5/K)}{\Gamma^2(3/K)}.$$

Последний можно оценить по экспериментальным данным.

Построим две теоретические модели распределений вида (4): одну с параметрами $\sigma_1(M_{21})$ и $K(Э_1)$, оцененными по всем экспериментальным данным; другую с параметрами $\sigma_0(M_{20})$ и $K(Э_0)$, оцененными без далеко отстоящего наблюдения.

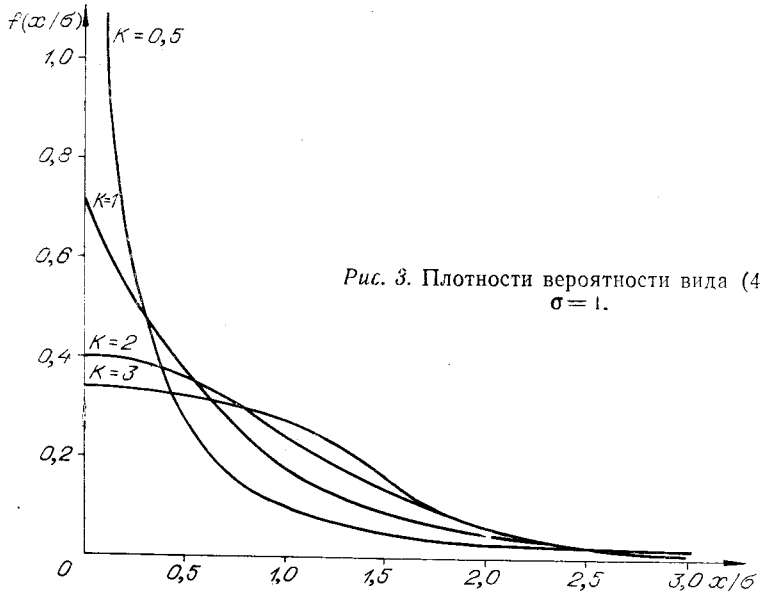

Рис. 3. Плотности вероятности вида (4) при $\sigma=1$.

При помощи критерия (1) проверим, какая из них ближе к распределению экспериментальных данных, имеющему далеко отстоящие наблюдения. Для этого разобьем сумму квадратов в (1) на два члена:

$$\sum \frac{(n_i - \nu_i)^2}{\nu_i} = \sum^{|x|/\sigma \leqslant 1} \frac{(n_i - \nu_i)^2}{\nu_i} + \sum^{|x|/\sigma > 1} \frac{(n_i - \nu_i)^2}{\nu_i}.$$

Первый член соответствует области случайных величин $|x|/\sigma \leqslant 1$, а второй — $|x|/\sigma > 1$.

В области $|x|/\sigma > 1$ распределение экспериментальных данных имеет «провал», т. е. участок с частостью n_i, равной нулю. Перед участком нулевой частости гистограмма имеет более крутой спад, чем в случаях отсутствия «провала». В то же время теоретическое распределение вида (4) обладает на участке «провала» гистограммы конечными значениями ν_i. Поэтому ближе к гистограмме экспериментальных данных будет в этой области та из теоретических кривых, которая перед провалом имеет большую крутизну спада, а на участке провала обладает меньшей плотностью. Такими свойствами обладает теоретическое распределение с параметрами $\sigma_0(M_{20})$ и $K(Э_0)$, если $\hat{Э}_0 < Э_1$ (см. рис. 3).

В области значений аргумента $|x|/\sigma \leqslant 1$ отношения $(n_i - \nu_i)^2/\nu_i$ при уменьшении оценки Э также уменьшаются. Это определяется следующими причинами:

1) эффективность оценок M_2 и Э увеличивается с уменьшением эксцесса;

2) модель, построенная по параметрам, рассчитанным только для компактной части распределения экспериментальных данных, лучше отражает ее специфику, так как вершина и склоны теоретической кривой плотности распределения меньше отклоняются от гистограммы экспериментальных данных и поэтому значение χ^2 уменьшается.

Таким образом, распределение вида (4), построенное по параметрам $\sigma(M_{20})$ и $K(Э_0)$, вычисленным без далеко отстоящего наблюдения, ближе к распределению экспериментальных данных.

Изложенное позволяет рекомендовать упрощенный метод исключения резко выделяющегося наблюдения с одновременным определением теоретического распределения с наилучшими параметрами по минимуму оценки эксцесса. Он состоит в следующем.

1. По качественным признакам выявляют наличие далеко отстоящего наблюдения. Такими признаками являются:

а) при сгруппированных данных наличие провала с нулевой частостью между компактной частью и далеко отстоящим наблюдением шириной более трех интервалов группирования;

б) при несгруппированных данных справедливость неравенства

$$|x_{n+1} - \bar{x}_0| \geqslant |x_n - \bar{x}_0| 1{,}5.$$

2. Вычисляют оценки эксцесса $Э_1$ и $\hat{Э}_0$. Если $\hat{Э}_0 < Э_1$, то наблюдение x_{n+1} можно отбросить и в качестве теоретического распределения принять модель (4), построенную по оценкам моментов $\hat{M}_{20}$ и $\hat{Э}_0$.

Для качественного решения вопроса о наличии далеко отстоящего наблюдения можно также воспользоваться графиком рис. 2. При этом можно отбросить наблюдения, для которых t_0 находятся в зоне А. Из рисунка видно, что принятие решения будет зависеть от оценки эксцесса. При равномерном распределении ($Э = 1{,}8$) отбросить можно наблюдение с $t_0 > 2$, при нормальном ($Э = 3$) — наблюдение с $t_0 > 3{,}5$, а при двойном экспоненциальном распределении ($Э = 6$) — только наблюдение с $t_0 > 5$. Наличие зоны неопределенности объясняется тем, что при расчетах мы используем не дисперсию и эксцесс, а их оценки.

Пример. На рис. 1 показана гистограмма распределения погрешностей измерения напряжения с частотой 50 Гц при помощи электрон-

ных вольтметров типа В3—2А в диапазоне 0÷30 мВ при номинальном значении измеряемого напряжения 5 мВ. Всего имеется $n=112$ наблюдений, одно из которых с погрешностью +13% далеко отстоит от остальных, образующих компактную часть распределения.

Если учитывать все экспериментальные данные, то получим следующие оценки числовых характеристик: $\bar{x}_1=-0{,}5$ мВ; $M_{21}=6{,}9$ мВ^2; $A_1=1{,}0$; $Э_1=7{,}4$.

Распределение с такими числовыми характеристиками следует считать асимметричным, и теоретическую модель его строить по формуле для асимметричных распределений, приведенной в [5]. В результате вычислений, проведенных по формуле (1), получается $\sum \bar{\chi}_i^2 = 20{,}61$ при числе степеней свободы $R=5$. Отсюда среднее значение $\bar{\chi}^2=4{,}1$, а вероятность $P=0{,}1\%$.

Если отбросить далеко отстоящее наблюдение, то оценки числовых характеристик будут равны:

$$\bar{x}_0=-0{,}6 \text{ мВ}; \ \hat{M}_{20}=5{,}4 \text{ мВ}^2; \ \hat{A}_0=-0{,}05; \ \hat{Э}_0=3{,}65.$$

Такое распределение экспериментальных данных следует признать симметричным и для построения модели воспользоваться формулой (4). При вычислениях по формуле (1) получается $\sum \chi_i^2=4{,}52$ при числе степеней свободы $R=6$; $\bar{\chi}^2=0{,}75$; $P=60\%$. Поэтому вторая модель более близка к распределению экспериментальных данных, то же самое показывают и оценки эксцесса.

ВЫВОДЫ

1. Вопрос о признании наблюдения далеко отстоящим должен решаться одновременно с выбором теоретической модели типа (4), наиболее близкой в смысле критерия минимума $\bar{\chi}^2$ или минимума эксцесса к распределению всех экспериментальных данных.

2. В тех случаях, когда совокупность экспериментальных данных состоит из компактной части и отдаленного наблюдения, это наблюдение не следует учитывать при расчетах параметров теоретической модели, если при его отбрасывании оценка эксцесса уменьшается.

ЛИТЕРАТУРА

1. **Большев Л. Н., Смирнов Н. В.** Таблицы математической статистики. М., «Наука», 1965. 464 с.
2. **Смирнов Н. В., Дунин-Барковский И. В.** Курс теории вероятностей и математической статистики. М., «Наука», 1969. 511 с.
3. **Абезгауз Г. Г., Тронь А. П., Копенкин Ю. Н., Коровина И. А.** Справочник по вероятностным расчетам. М., Воениздат, 1970. 535 с.
4. **Новицкий П. В., Назаров И. А., Иванова В. Я., Кондрашкова Г. А.** Сравнение оценок погрешностей измерений по энтропийному, среднеквадратическому и предельному значениям.— «Измерительная техника», 1966, № 9, с. 20—24.
5. **Резник К. А.** Об одной модели распределения погрешностей ансамблей измерительных приборов.— «Автометрия», 1970, № 5, с. 46—49.
6. **Резник К. А.** Использование свойств одной модели распределения при нормировании погрешностей средств измерений.— «Автометрия», 1972, № 1, с. 19—23.
7. **Крамер Г.** Математические методы статистики. М., «Мир», 1975. 647 с.

Поступила в редакцию 25 июля 1975 г.;
окончательный вариант — 7 июля 1976 г.