ИССЛЕДОВАНИЕ ГИСТОГРАММ ГЕОЛОГИЧЕСКИХ ПРИЗНАКОВ КОМПЬЮТЕРНЫМ МОДЕЛИРОВАНИЕМ

Д. г.-м. н.
Ю. А. Ткачев
tkachev@geo.komisc.ru

Гистограмма, как известно, представляет собой столбчатый график распределения частот по интервалам значений признака. Анализ гистограмм традиционно считается среди геологов наглядным и информативным методом решения геологических задач. Даже простое сведение цифровой информации в рисунок приносит большую пользу. Гистограмма позволяет единым взглядом охватить всю выборку и составить представление о распределении изучаемой величины, чего нельзя сделать даже внимательным изучением таблиц исходных значений. Анализ гистограмм позволяет проверять геологические гипотезы, сформулированные на языке статистики. Одним из наиболее простых действий приемов анализа гистограмм является проверка гипотезы о том, что распределение данных подчиняется предполагаемому, например нормальному (гауссовскому), закону. Техника проверки заключается в сравнении эмпирических ($n_э$) и теоретических ($n_т$) частот. Расчет теоретических частот производится при условии, что выборка данных получена из генеральной совокупности с предполагаемым, например гауссовским, распределением. Сравнить эти частоты с эмпирическими можно с помощью критерия "согласия" хи-квадрат:

$$\chi^2 = \sum \frac{(n_т - n_э)^2}{n_т}.$$

В зависимости от полученного значения χ^2 принимают решение. Если значение χ^2 превышает критическое,

$$\chi^2_{f,\alpha},$$

где f — число степеней свободы, $f = k - \nu$, где k — число интервалов гистограммы, ν — число параметров "подозреваемого" закона, оцененных по эмпирическим данным, α — уровень значимости (так называемая ошибка первого рода — вероятность отклонения верной гипотезы), то гипотезу отклоняют, если значение χ^2 меньше критического — не отклоняют. Гипотезу о принадлежности выборки данному закону называют *нулевой* H_0. В данном случае

$$H_0 : f(x) = G,$$

где G — гауссовское распределение. Рассмотренная нулевая гипотеза является *простой* (независимо от того, сложен или несложен предполагаемый закон), потому что теоретические частоты по этому закону являются вполне определенными, вычислимыми.

Как это ни парадоксально, проверить "менее сильные" гипотезы оказывается значительно сложнее. Под менее сильной гипотезой мы подразумеваем гипотезу более общего, менее притязательного характера, как правило, менее специальную. Например, менее сильной гипотезой является гипотеза, согласно которой распределение симметрично относительно того или иного интервала гистограммы. Менее сильные гипотезы мы называем сложными, так как для них могут существовать много вариантов теоретических частот.

Если верна более сильная гипотеза, то подавно верна и менее сильная, но не наоборот. Например, из гауссовского закона следует, что распределение симметрично, из симметричности гистограммы не следует, что распределение подчиняется гауссовскому закону. Образование планеты Земля из планетезималей — гипотеза менее сильная, чем гипотеза образования её из твердых частиц размером до 1 мм.

Менее сильной гипотезой, чем принадлежность распределения гауссовскому закону, является гипотеза, что распределение, представленное на гистограмме, одномодально. Имеющимися средствами математической статистики в общем случае такой вопрос разрешить сложнее, чем установить, подчиняется ли распределение указанному наперед закону. Если бы у нас была только одна определенная альтернатива многомодальному распределению (например, нормальное распределение), задача решалась бы просто: необходимо было бы проверить гипотезу о нормальности и в случае её отклонения считать, что наше распределение би- или полимодальное. Однако существует **бесчисленное** множество теоретических вариантов одномодальных распределений, не подчиняющихся гауссовскому закону. Наш вариант проверки может незаконно их отклонить.

В наиболее общем виде задача, поставленная в статье, формулируется следующим образом: *на заданном уровне значимости проверить гипотезу о том, что представленное на гистограмме распределение является мономодальным.* Поскольку мономодальных распределений бесконечное количество, проверяемая гипотеза оказывается сложной:

$$H_0 : \{(F_э = F_1) \vee (F_э = F_2) \vee \ldots$$
$$\ldots (F_э = F_i) \ldots \vee (F_э = F_n)\},$$
$$n \to \infty \qquad (1)$$

Точнее, следовало бы в записи (1) писать не $F_э = F_i$, а "эмпирическое распределение $F_э$ есть распределение F_1 или F_2, или ... F_n". Еще точнее нулевую гипотезу необходимо было бы сформулировать так: "выборка, представленная гистограммой, является выборкой из одномодального распределения F_1 или F_2, или … F_n". Именно в этом смысле мы будем понимать запись (1).

Проверить бесконечное число гипотез, образующих в совокупности сложную слабую гипотезу, невозможно, и мы вынуждены сделать ряд упрощений.

1. Распределения: а) проверяемое эмпирическое и б) теоретические одномодальные, будем считать *дискретными*. Конкретнее, случайная величина в них может принимать k значений — по числу интервалов проверяемой гистограммы.

2. Частоты значений в каждом теоретическом распределении (т.е. фактически сами распределения) определяются числом элементов, размещенных в i-том интервале. Общее число элементов при таком расчете равно объему

выборки n, т.е. $h_i = \frac{n_i}{n}$, причем n элементов должны быть размещены в k интервалах-ячейках так, чтобы образовалась мономодальная гистограмма. Каждое такое размещение создает одно теоретическое мономодальное распределение.

Общее число размещений n элементов по k ячейкам (в каждой ячейке может быть n_i элементов при условии $n \geq n_i \geq 0$ и $\sum n_i = n$) рассмотрим на следующих примерах.

Вариант 1. Все элементы сосредоточены в одном интервале, остальные интервалы пусты: $M_1 = k$, где M_1 — число простых гипотез. Вариант возможен при любом соотношении n и k. Гипотезы из варианта 1 проверены быть не могут, так как число степеней свободы равно нулю: $k = 1, f = k - 1$ (одна связь наложена: общая численность теоретических частот равна таковой в эмпирической гистограмме).

Вариант 2. В одном интервале сосредоточено $n - 1$ элемент, один элемент — в любом другом:

$M_2 = k \cdot (k - 1)$.

Вариант 3. В одном интервале сосредоточено $n - 2$ элемента, два элемента сосредоточено в остальных, т. е. 1-й элемент попадает в $k - 1$ интервал, 2-й элемент — в $k - 2$ оставшихся, т. е.

$$M_3^1 = k \cdot (k-1),$$

$$M_3^2 = k \cdot (k-1)(k-2).$$

Вариант 4. Аналогично:

$$M_4^1 = k \cdot (k-1),$$

$$M_4^2 = k \cdot (k-1)(k-2),$$

$$M_4^3 = k \cdot (k-1)(k-2)(k-3).$$

Вариант 5. По тому же принципу устанавливаем:

$$M_5^1 = k \cdot (k-1),$$

$$M_5^2 = k \cdot (k-1)(k-2),$$

$$M_5^3 = k \cdot (k-1)(k-2)(k-3),$$

$$M_5^4 = k \cdot (k-1)(k-2)(k-3)(k-4),$$

и т. д.

Общее число таких теоретических дискретных распределений будет равно сумме

$$M_1 + M_2 + M_3^1 + M_3^2 + M_4^1 + M_4^2 + + M_4^3 + M_5^1 ... + M_5^4.$$

Эта сумма равна числу разбиений R_n целого числа n на k целых положительных слагаемых s_i, $0 \leq s_i \leq n$. Это число конечно (хотя при больших n и велико), что в принципе позволяет запрограммировать моделирование этих разбиений. В процессе моделирования их можно посчитать, так что при таком подходе отсутствие формулы числа разбиений не препятствует решению задачи. Из разбиений необходимо отобрать мономодальные распределения числом $R_{n,\text{моно}}$ по принципу неуменьшения частот к некоторому i-тому интервалу и неувеличения их после этого интервала.

Таким образом получают все дискретные мономодальные распределения, возможные при определенных объемах выборки n и числа интервалов k. В общем случае каждое из этих распределений имеет одинаковую вероятность появления, равную $P_i = R_{n,\text{моно}}/R_n = const$. При некоторых специальных условиях можно рассматривать и неравные вероятности моделируемых распределений.

Специально рассмотрим вопрос о том, как понимать проверку составной гипотезы типа (1) на заданном уровне значимости α. Здесь может быть две трактовки. Первая из них заключается в том, что H_0 не отклоняется, если из всего многообразия мономодальных распределений $\{F_i\}$ найдется хотя бы одно, не отклоняемое на уровне значимости α, или, что то же самое, имеется хотя бы одно F_i, принимаемое с доверительной вероятностью $P = 1 - \alpha$. При таком понимании эмпирическое распределение не противоречит хотя бы одному из предложенных мономодальных распределений. Рассмотрим эту трактовку более подробно.

Если существует такое мономодальное распределение, которому не противоречит наше эмпирическое на уровне значимости α, то гипотеза мономодальности не отклоняется. Следовательно, принимать гипотезу би- или полимодальности нет оснований и имеющиеся "провалы" на эмпирической гистограмме следует считать несущественными. На языке математической логики это запишется следующим образом:

если

$$\exists (H_{oi})\{P(H_{oi} = true) \geq 1 - \alpha_{кр}\},$$

$$i = 1..M, \qquad (2)$$

где P — вероятность истинности i-той гипотезы H_{oi}, $\alpha_{кр}$ — принятый уровень значимости, i — номер теоретического мономодального распределения, M — общее число таких распределений, то распределение $F_э$ мономодально.

По другому условие (2) можно записать следующим образом:

$$\exists \alpha_i \{\alpha_i > \alpha_{кр}\},$$

или

$$\exists i \{\chi_i^2 < \chi_{\alpha_{кр}}^2\}$$

Однако такая трактовка, на наш взгляд, не вполне соответствует существу задачи.

В другом, более приемлемом на наш взгляд варианте, гипотеза мономодальности может быть отклонена также и в том случае, если условие (2) по отдельности не выполняется ни для одного мономодального распределения, но её отклонение (или принятие) делается по совокупности полученных значений уровня значимости $\{\alpha_i\}$, .

Действительно, для неотклонения (принятия) гипотезы мономодальности необходимо, чтобы доверительная вероятность такого решения P была бы больше критической, т. е.

$P > P_{кр} = 1 - \alpha_{кр}$;

Если проверяемая гипотеза составная, то необходимо, чтобы

$P(A_1 + A_2 + A_3 + ... A_n) > P_{кр}$,

где A_i – событие, что гипотеза H_{oi} может быть принята с доверительной вероятностью $P_i = 1 - \alpha_i$. Для простоты рассмотрения ограничимся случаем, в котором гипотеза мономодальности состоит всего из двух простых гипотез H_{o1} и H_{o2}. Тогда

$P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1A_2)$,

так как события A_1 и A_2 — совместимые (наше эмпирическое распределение может не противоречить сразу нескольким мономодальным "теоретическим" распределениям).

Проиллюстрируем вторую трактовку проверки сложной составной гипотезы несколькими примерами, сведенными в табл. 1.

Проверка каждой из трех гистограмм проводилась по критерию χ^2. Результаты приведены в первых строках для каждой простой гипотезы H_{o1}, H_{o2} и H_{o3}. В первом эмпирическом распределении доверительная вероятность принятия гипотез H_{o1} и H_{o2} мономодальности не достигает требуемого уровня. Однако по совокупности проверок обеих гипотез гипотеза мономодальности принимается (срабатывает вторая трактовка!).

Таблица 1

Пример последовательной проверки гипотезы мономодальности
(критическое значение $\alpha = 0.1$ (10%), число степеней свободы $f = 10$)

Номер теоретич. распределения	Гистограмма 1*			Гистограмма 2			Гистограмма 3		
1	χ^2	α	P	χ^2	α	P	χ^2	α	P
	13.1	0.25	0.75	25.2	0.05	0.95	9.34	0.5	0.5
	$P_1 = 0.75 < P_{кр} = 0.9$, гипотеза H_{o1} не принимается			$P_1 = 0.95 > P_{кр} = 0.90$, гипотеза мономодальности принимается на основании проверки H_{o1}			$P_1 = 0.5 < P_{кр} = 0.9$, гипотеза мономодальности на основании проверки H_{o1} не принимается		
2	13.4	0.20	0.80	Проверка гипотезы 2 не требуется			8.30	0.6	0.4
	$P_2 = 0.8 < P_{кр} = 0.9$, гипотеза H_{o2} не принимается						$P_2 = 0.4 < P_{кр} = 0.9$, гипотеза мономодальности на основании проверки H_{o2} не принимается		
1 + 2	По совокупности проверок гипотез H_{o1} и H_{o2}: $P_{1,2}(A_1 + A_2) = 0.75 + 0.80 - 0.75 \cdot 0.80 = 1.55 - 0.60 = 0.95$ $P_{1,2} = 0.95 > P_{кр} = 0.90$; гипотеза мономодальности принимается по совокупности проверок двух простых гипотез			Проверка по совокупности не требуется			По совокупности проверок гипотез H_{o1} и H_{o2}: $P_{1,2}(A_1 + A_2) = 0.5 + 0.4 - 0.4 \cdot 0.5 = 0.9 - 0.2 = 0.7$ $P_{1,2} = 0.7 < P_{кр} = 0.90$; гипотеза мономодальности по совокупности проверок гипотез H_{o1} и H_{o2} не принимается		
3	Проверка гипотезы 3 уже не требуется			Проверка гипотезы 3 не требуется			9.30	0.5	0.5
							$P_3 = 0.5 < P_{кр} = 0.9$, гипотеза мономодальности на основании проверки H_{o3} не принимается		
1 + 2 + 3	—			—			По совокупности проверок гипотез H_{o1}, H_{o2} и H_{o3}: $P_{1,2,3}(A_1 + A_2 + A_3) = 0.7 + 0.5 - 0.7 \cdot 0.5 = 1.2 - 0.35 = 0.85$; $P_{1,2,3} = 0.85 < P_{кр} = 0.9$; гипотеза мономодальности по совокупности проверок гипотез H_{o1}, H_{o2}, H_{o3} не принимается		

* Модельные примеры гистограмм

Во втором эмпирическом распределении мономодальность принимается уже по результатам проверки первой гипотезы H_{o1}, поэтому дальнейшие проверки не требуются: они никаким образом не могут уменьшить доверительную вероятность. Тем не менее проверки других имеющихся простых гипотез, составляющих сложную нулевую гипотезу, если таковые имеются, могут быть продолжены для установления действительного значения доверительной вероятности, что в некоторых случаях бывает полезным или необходимым.

В третьем эмпирическом распределении гипотеза мономодальности не принята ни по результатам проверки отдельных гипотез H_{o1}, H_{o2}, H_{o3}, ни по их совокупности. Если других простых гипотез мономодальности нет, то результат проверки сложной гипотезы следует считать окончательным. Если существуют другие теоретические мономодальные распределения кроме испытанных H_{o1}, H_{o2} и H_{o3}, проверка должна быть продолжена либо до исчерпания H_{oi}, либо до достижения критического значения доверительной вероятности.

В связи с изложенными выше процедурами проверки сложных гипотез возникает следующая проблема. Она заключается в решении вопроса о том, следует ли учитывать вероятности появления в данной предметной области "образцовых" для сравнения мономодальных распределений, с которыми сравнивается наше эмпирическое. При проверке простой гипотезы, например гипотезы принадлежности эмпирического распределения гауссовскому, также можно задаться вопросом, какова априорная вероятность встречаемости этого распределения в данной предметной области. Если она равна нулю, то независимо от результатов применения критерия согласия необходимо констатировать, что эмпирическое распределение в данной ситуации не может быть гауссовским. И наоборот, если гауссовское распределение здесь единственно возможное, то независимо от проверки принадлежности с необходимостью вытекает, что эмпирическое распределение – гауссовское. В общем случае

$$P = P_д \cdot P_р, \qquad (5)$$

где P – вероятность истинности принимаемой гипотезы, $P_д$ – доверительная вероятность, полученная в результате проверки гипотезы (или установленная заранее), $P_р$ – априорная вероятность распространения данного "теоретичес-

кого" распределения, испытуемого в качестве нулевой гипотезы.

В обычной практике проверки гипотез о законах распределений вероятность $P_р$ не устанавливается и не учитывается, что соответствует условиям полной неизвестности априорных вероятностей распространенности видов распределений, когда для всех подбираемых для сравнения законов принимается $P_р = const$. Мы можем поступить так же, но в нашем случае объектов для сравнения с ними эмпирического распределения может быть много тысяч, и вопрос о том, какие из них вероятны в природе, не тривиален. Без учета распространенности тех или иных распределений описанная выше процедура должна квалифицироваться как подгонка. Ввод в такие процедуры априорных вероятностей представляется нам совершенно необходимым элементом.

Технические трудности применения критерия согласия χ^2 к дискретным распределениям выражаются в том, что во многих теоретических мономодальных распределениях часть интервалов будет иметь нулевые частоты, что недопустимо, так как они в формуле для расчета χ^2 появляются как в числителе, так и в знаменателе. Для преодоления этой трудности "нулевые" интервалы объединяются с соседними, чтобы ни в одном не оказалось нулевых частот. При этом аналогичным образом объединяются с суммированием частот те же интервалы эмпирической гистограммы. После такой фильтрации в эмпирической гистограмме могут оставаться интервалы с частотами меньше пяти, что при проверке допускать не рекомендуется (по другим источникам частоты должны быть не менее трех). Снова применяем фильтр с объединением интервалов эмпирической гистограммы, а также соответствующих интервалов теоретического распределения. После такой подготовки используем критерий χ^2.

Статистические таблицы, содержащие значения хи-квадрат распределения для различных степеней свободы и критические значения χ^2 для разных уровней значимости α, оказались непригодными для решения нашей задачи. Таблицы сделаны для дискретных значений уровня значимости, а нам необходимы точные значения α для получаемых в результате применения формулы (1) значений χ^2. В программах, реализующих описываемую методику, необходимо было предусмотреть способ расчета α для заданных степеней свободы и получаемых значений χ^2. Для этого в свою очередь необходимо было разработать ряд процедур вычисления специальных математических функций, и в их числе гамма-функцию — аналога факториала для дробного аргумента.

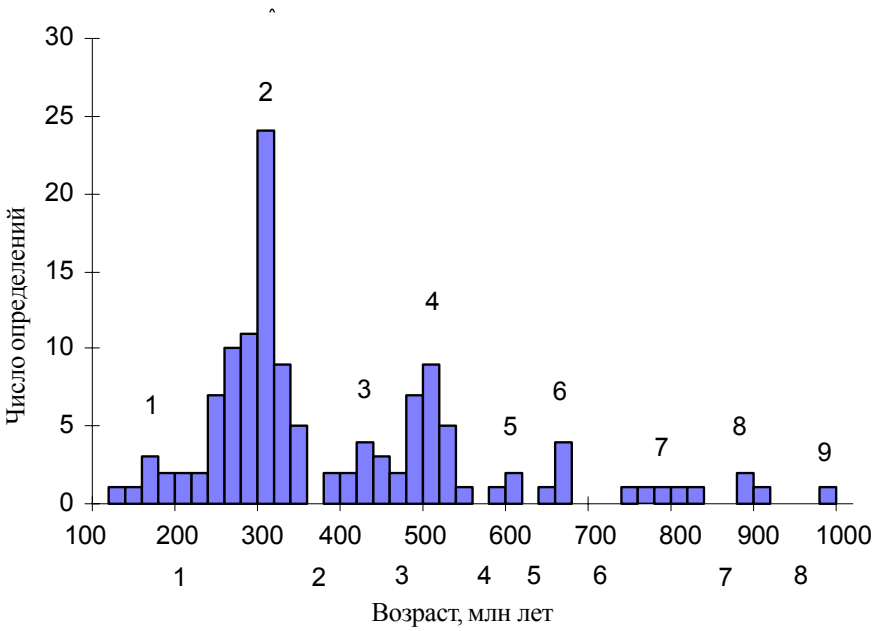
Разработанный нами программный комплекс на языке Паскаль-7 содержит ... программных единиц общим объемом ... операторов. Он способен обрабатывать гистограммы, содержащие до 25 интервалов с объемом выборки до 300. Обработка больших объемов выборки сильно удлиняет время счета, и без того немалое.

Методика испытана на небольших искусственных примерах, а также на реальном материале, любезно предоставленном Н. П. Юшкиным, которому автор обязан также постановкой задачи именно в широком плане, без предположений о виде теоретического распределения, как эталона для сравнения с эмпирическим. Исходные данные были заданы в виде гистограммы (см. рисунок). Необходимо было установить, какие из "провалов" гистограммы существенны *без априорных предположений о виде распределения* случайных величин в целой гистограмме и в отдельных её частях.

При анализе сложных гистограмм, содержащих несколько мод и разделяющих их "провалов", предусмотрена такая последовательность действий. Сначала рассчитывается значение χ^2 и α в целом для всей гистограммы. Если гипотеза мономодальности не отклоняется, анализ можно закончить. Если гистограмма разделяется на две части по интервалу самого существенного провала, то анализ продолжается аналогичным образом для каждой части. На практике "для надежности", независимо от результатов проверки общей гистограммы или её части с "провалом", анализ ведется до конца.

Для ускорения анализа можно воспользоваться вариантом программы, в котором не предусматривается моделирование мономодальных распределений, а используется ручной ввод частот мономодальной гистограммы, на глаз наименее отличающейся от эмпирической. Если при этом χ^2_1 не превысит критического значения α, эмпирическая гистограмма считается мономодальной. В противоположном случае частотами "теоретической" мономодальной гистограммы немного варьируют, но в пределах, оставляющих её мономодальной, анализ повторяется и ведется проверка по описанным схемам первой и второй трактовок проверки сложных гипотез.

В реальном примере исходные данные гистограммы (см. рисунок) представляют собой определения возраста цирконов из долины р. Оби изотопным методом (или по трекам осколков деления?). В этой гистограмме насчитыва-



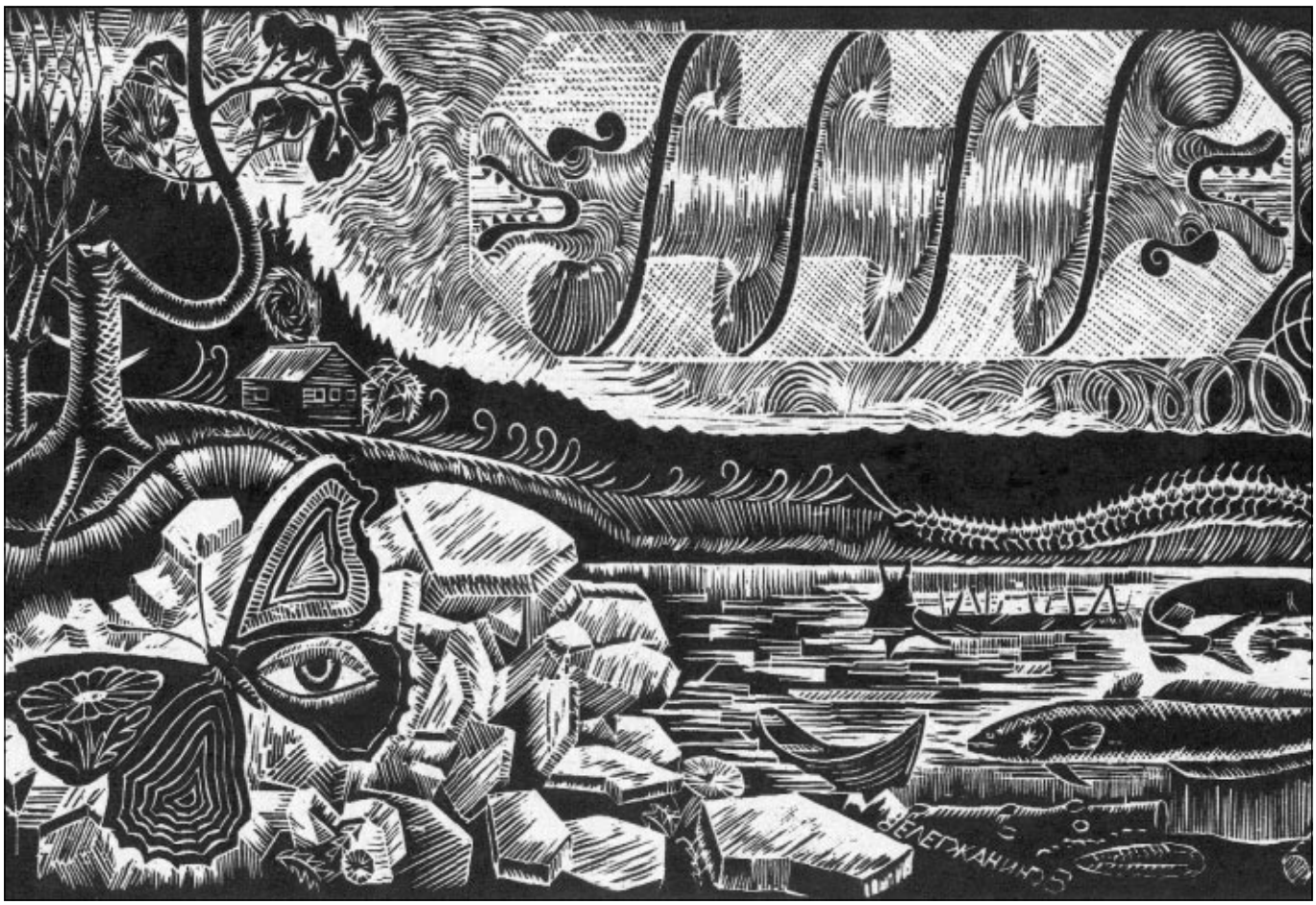
Исходная гистограмма распределения цирконов из долины р. Оби по возрасту, млн лет. Цифры сверху – номера модальных интервалов, цифры снизу – номера провальных интервалов. Результаты анализа приведены в табл. 2

Таблица 2

Результаты проверки на мономодальность исходной гистограммы и ее частей (критический уровень значимости $a_{кр} = 0.1$)

Анализируемая часть гистограммы	Число интервалов k	Число данных n	Число степеней свободы f	Значение χ^2	Уровень значимости α	Вывод
Левая от провала 4	23	112	13	55.2	$0.4 \cdot 10^{-6}$	Провал 2 существен
Левая от начала до провала 2	13	77	7	1.03	0.996	Провал 1 несуществен
От провала 2 до провала 4	9	35	5	2.41	0.75	Провал 3 несуществен
От провала 4 до провала 6	6	8	1	0.67	0.42	Провал 5 несуществен
От провала 4 до правого конца гистограммы	21	17	2	10.3	0.005	Провал 6 существен
От провала 6 до провала 8	14	8	3	10.0	0.002	Провал 7 существен
От провала 7 до правого конца гистограммы	7	4	0	—	—	О существенности провала 8 судить невозможно ввиду отсутствия степеней свободы

ется девять мод. Визуально наиболее контрастным представляется провал между модами 2 и 3. Расчеты это подтверждают. Поэтому первым шагом является расчленение гистограммы по провалу между модами 4 и 5 с целью анализа левой части гистограммы. Результат анализа этой части гистограммы на мономодальность приведен в первой строке табл. 2: она не мономодальна, и вероятность ошибки этого вывода не превысит $0.4 \cdot 10^{-6}$, что ничтожно мало. Дальнейший ход анализа и его результаты приведены в этой же таблице.



Гравюра О. Велегжанинова “Бабочка на льду”