В. Е. ЯГУР, В. В. НЕШИТОЙ, И. И. САЛИВОН

НОВЫЙ ПОДХОД К СТАТИСТИЧЕСКОМУ АНАЛИЗУ БИОМЕДИЦИНСКИХ ДАННЫХ

Белорусский государственный медицинский университет,
Белорусский государственный университет культуры
и искусств, Институт истории НАН Беларуси

С помощью методов, разработанных в рамках теории обобщенных распределений, исследованы возможности оценки параметров эмпирических распределений биомедицинских данных на примере анализа антропометрических признаков: длины тела, массы тела, индекса Кетле у мужчин в возрасте от 18 до 36 лет (n=316). Эмпирические распределения указанных признаков хорошо аппроксимируются теоретическим распределением 3-го типа. Параметры теоретической кривой С (мода), А и В (точки перегиба) позволяют математически строго выделить центральный (от А до В) и периферические диапазоны (<A и >B), интерпретируя центральный диапазон как популяционную норму. Вычисленные диапазоны были использованы для разработки системы полуколичественного (в баллах) способа соматотипирования мужчин в возрасте от 18 до 36 лет. Показана возможность наглядной визуализации аппроксимирующих распределений в виде точек на плоскости «асимметрия/эксцесс», что позволяет оценить удаленность этих точек (популяций) друг от друга и от известных теоретических непрерывных распределений (нормальное, логнормальное, Лапласа, Коши, Вейбулла и др.).

Ключевые слова: *биомедицинские данные, антропометрические признаки, эмпирическое распределение, теоретическое распределение, кривая аппроксимации.*

При статистическом анализе биомедицинских данных весьма важной является проблема выбора наиболее подходящего закона распределения для описания группированного вариационного ряда. Выборочные характеристики — среднее арифметическое значение выборки и традиционно используемые показатели ее варьирования (дисперсия, стандартное отклонение, коэффициент вариации) — не содержат полной информации о законе распределения генеральной совокупности, из которой выборка взята.

Далеко не всегда можно судить о законе распределения по эмпирической вариационной кривой, поскольку на ней сказывается влияние различных определенных и неопределенных случайных причин. Между тем знание закона распределения биомедицинских признаков защищает от возможных ошибок в оценке генеральных параметров на основании выборочных показателей, а также позволяет использовать наиболее адекватные и мощные методы их статистического анализа [2].

Чаще всего результаты измерений количественных характеристик биомедицинских объектов в интервальной шкале или шкале отношений существенно отличаются от нормального (гауссово) распределения [8, 9, 11].

По данным ряда исследований, лишь 10—20% распределений количественных признаков, встречающихся в биомедицинских исследованиях, являются приближенно нормальными. А это означает, что применение параметрических методов статистики (t-критерий Стьюдента, дисперсионный, регрессионный, факторный анализы) в большинстве случаев по меньшей мере не является обоснованным. Для указанных видов анализа данных параметрические методы могут применяться лишь в том случае, когда все анализируемые одновременно признаки имеют нормальное распределение, иначе должны использоваться непараметрические методы статистики [9] или иные подходы, о которых пойдет речь в статье.

Для достаточно надежного установления нормальности требуется либо весьма большое число наблюдений (порядка нескольких тысяч), что крайне редко встречается в медико-биологических исследованиях, либо проверка статистических гипотез независимости и одинаковой распределенности с помощью соответствующих непараметрических критериев согласия — Смирнова, Колмогорова, Лемана—Розеблатта (омега-квадрат), критерий согласия Пирсона χ^2 и др. Условия применимости t-критерия Стьюдента весьма ограничены. Он может использоваться для проверки нулевой гипотезы (H_0) об однородности математических ожиданий в двух выборках при выполнении следующих условий: 1) результаты наблюдений имеют нормальные распределения; 2) дисперсии в двух выборках совпадают [10].

В прикладной статистике разработаны два подхода к обработке исходных данных — детерминированный и модельно-вероятностный. При детерминированном подходе данные анализируют сами по себе, без переноса выводов с конкретной ситуации на другие, ей подобные, и нет нужды в знании характера распределения данных, но, к сожалению, невозможно оценить погрешность рассчитанных показателей.

Цель научного исследования заключается в выявлении некоторых закономерностей на выборке и дальнейшей экстраполяции полученных результатов на всю генеральную совокупность (популяция), из которой получена исследуемая выборка. Для этого и используется модельно-вероятностный подход, согласно которому в основе алгоритма расчета лежит вероятностная модель порождения данных, наиболее распространенной из которых является модель случайной выборки.

Существуют различные параметрические семейства распределений числовых случайных величин (семейства нормальных распределений, логарифмически нормальных, экспоненциальных, гамма-распределений). Все они зависят от одного или нескольких параметров, знания которых достаточно для полного описания распределения.

Другой подход заключается в применении методов преобразования (трансформации) данных имеющегося эмпирического распределения в нормальное с помощью логарифмирования (без сдвига и со сдвигом), извлечения квадратного корня, обратного преобразования и др. В ряде случаев такая трансформация может оказаться удачной в плане приближения эмпирических данных к нормальному распределению, но следует помнить, что преобразование данных ведет к изменению единиц их измерения и потере ими своего физического смысла, поэтому результаты анализа преобразованных данных трудно оценивать [1].

Существует и третий подход определения характера эмпирического распределения, который предложен доктором технических наук В. В. Нешитым [3—7].

Им разработана теория обобщенных (универсальных) распределений, которая включает три системы непрерывных распределений, систему дискретных распределений, взаимосвязанную с системой кривых роста новых событий, методы установления типа выравнивающей кривой и нахождения оценок параметров по номограммам. Автор также предложил серию компьютерных программ для работы с указанными системами.

Обычно в качестве основной вероятностной модели используется нормальный закон распределения, но с его помощью чаще всего не удается с достаточной точностью выровнять (аппроксимировать) эмпирические распределения. Кроме того, существуют случайные величины, распределение которых в принципе не может быть описано нормальным законом.

Для выравнивания большого разнообразия статистических распределений В. В. Нешитым построены три системы непрерывных распределений, каждая из которых задается одной, двумя или тремя плотностями.

Первая система непрерывных распределений используется для выравнивания и прогнозирования статистических распределений таких случайных величин, последующие значения которых образуются из предыдущих путем их изменения (сдвига) на некоторую постоянную величину. При этом форма кривой распределения не изменяется. Не изменяются также центральные моменты, но изменяется среднее значение и коэффициент вариации. Средние значения таких случайных величин растут во времени по линейному закону, но могут подчиняться другому закону роста. Частным случаем первой системы непрерывных распределений является нормальный закон.

Вторая система непрерывных распределений используется для выравнивания и прогнозирования статистических распределений таких неотрицательных случайных величин, последующие значения которых образуются из предыдущих путем их умножения на некоторую постоянную величину. При этом форма кривой распределения меняется, но неизменными остаются коэффициент вариации и параметры k, u. Средние значения логарифмов таких случайных величин изменяются во времени по линейному закону, а сами случайные величины — по показательному закону. Частным случаем второй системы непрерывных распределений является логарифмически нормальный закон.

Третья система непрерывных распределений используется для выравнивания и прогнозирования статистических распределений таких неотрицательных случайных величин, последующие значения которых образуются из предыдущих путем их возведения в некоторую степень. Средние значения двойных логарифмов таких величин изменяются во времени по линейному закону, а сами случайные величины — по двойному показательному закону. Частным случаем третьей системы непрерывных распределений является двойное логарифмически нормальное распределение, то есть случайная величина «lnlnY» распределена по нормальному закону.

Теория обобщенных распределений позволяет вычислять минимально необходимый объем выборки (n), при котором распределение среднего арифметического n независимых одинаково распределенных случайных величин можно считать нормальным. Кроме того, можно рассчитать доверительный интервал для среднего при заданной доверительной вероятности и объеме выборки при любом законе распределения случайной величины, а также рассчитать коэффициенты асимметрии и эксцесса.

Материал и методы

Рассмотрим возможности оценки параметров распределения биомедицинских данных на примере анализа антропометрических признаков: длины тела (ДТ), массы тела (МТ), индекса Кетле (ИК) у мужчин в возрасте от 18 до 36 лет (n=316). Данная выборка представляет собой фрагмент материалов по изучению индивидуальных и популяционных особенностей физического типа белорусов, собранных соавтором данной статьи (И. И. Саливон) во время комплексных экспедиций в 1975—1985 гг., организованных отделом антропологии и экологии Института искусствоведения, этнографии и фольклора НАН Беларуси в различных регионах Беларуси [12].

Избранные для анализа антропометрические признаки были аппроксимированы с помощью традиционного подхода — нормальным и логнормальным распределениями, а также с помощью второй системы непрерывных распределений по В. В. Нешитому, которая задается тремя обобщенными плотностями:

$$\left.\begin{aligned} p(t) &= Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1} \\ p(y) &= \frac{N\left(\ln y - l\right)^{k-1}}{y}\left[1-\alpha u(\ln y - l)\right]^{\frac{1}{u}-1} \\ p(y) &= \frac{N}{y}\left[1-\alpha u(\ln y - \overline{\ln y})^{2}\right]^{\frac{1}{u}-1} \end{aligned}\right\}$$

где N — нормирующий показатель; k, β, α, u, l — параметры.

Третья плотность включает логнормальный закон (при $u \to 0$).

В зависимости от значений параметров u, α, а также знака параметра β все распределения разделяются на типы (рис. 1).

Для плотности p(t) нормирующий множитель в зависимости от типа распределения задается формулами:

Типы I, I′: $$N = \frac{\beta(\alpha u)^k \Gamma(k+1/u)}{\Gamma(k)\Gamma(1/u)};$$

Типы II, II′: $$N = \frac{\beta\alpha^k}{\Gamma(k)};$$

Типы III—V: $$N = \frac{\beta(-\alpha u)^k \Gamma(1-1/u)}{\Gamma(k)\Gamma(1-1/u-k)}.$$

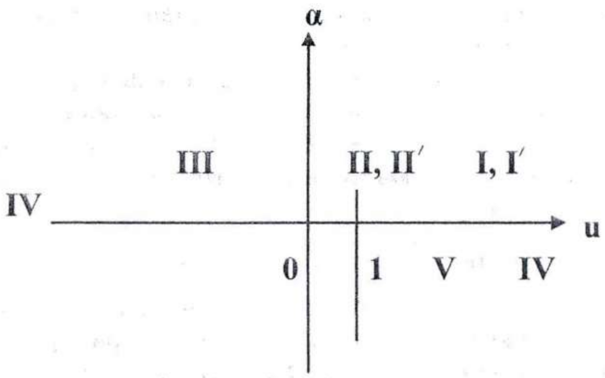

Рис. 1. Классификация типов распределений (типы со штрихом при β<0)

Для вычисления аппроксимирующего распределения и нахождения оценок параметров по статистическому интервальному ряду использован устойчивый метод [7].

По статистическим данным вычислены показатели ν_1^*, S_1^*, B^*, H^* по формулам:

$$\left.\begin{aligned} &\nu_1^* = \overline{\ln t} = \sum_{i=1}^{n} \ln t_i \frac{m_i}{M};\ \ S_1^* = \sum_{i=1}^{n} t_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} \\ &S_3^* = \sum_{i=1}^{n} t_i^3 \left(\frac{m_i}{M}\right)^4 \frac{1}{h^3};\ H^* = \frac{S_3^*}{(S_1^*)^3} \\ &B^* = \sum_{i=1}^{n} t_i \ln t_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} - \nu_1^* S_1^* \end{aligned}\right\}$$

Здесь $M = \sum_{i=1}^{n} m_i$ — сумма частот всех n интервалов, то есть объем выборки; h - ширина интервала; t_i — середина i-го интервала; B* - показатель асимметрии $(-1/4<B<1/4)$; H* - показатель островершинности $(1 < H^* < 2)$.

Показатели B и H зависят лишь от двух параметров — k, u. Это дает возможность построить в системе координат (B, H) бинарную сетку (номограмму), с помощью которой по статистическим показателям B^*, H^* можно установить тип аппроксимирующей кривой распределения и приближенно оценить показатели k, u. Два других показателя b, a (или произведение au) вычисляются по простым формулам [7].

Вычисление аппроксимирующих распределений по указанным выше формулам проведено с помощью компьютерной программы SNR2V08A, разработанной автором теории обобщенных распределений [7].

Подбор закона распределения для антропометрических признаков (ДТ, МТ, ИК), построение гистограмм их эмпирических распределений и аппроксимирующих кривых в соответствии с нормальным и логнормальным распределениями, а также рядом других теоретических распределений осуществляли с помощью программы STATISTICA 6.0. Оценку статистической значимости различий между фактическими и ожидаемыми значениями изученных признаков проводили с помощью критерия согласия Пирсона χ^2 [13].

Результаты и обсуждение

Традиционные расчеты средних значений (mean) и стандартных отклонений (SD) ДТ, МТ и ИК представлены в табл. 1. Кроме того, в этой же таблице приведены другие структурные средние: медиана (median), минимум (min), максимум (max), нижний (LQ_{25}) и верхний (UQ_{75}) квартили.

Гистограммы эмпирических распределений признаков ДТ, МТ, ИК, а также их аппроксимации с помощью кривой нормального закона распределения представлены на рис. 2—4, аппроксимация признака ИК с помощью кривой логнормального закона распределения — на рис. 5.

Гистограмма эмпирического распределения признака ДТ соответствует теоретическому нормальному закону (χ^2=10,1; P=0,53).

Признак МТ не аппроксимируется нормальным распределением (χ^2=20,5; P=0,002). Это означает, что нулевая гипотеза (H_0) о соответствии эмпирического распределения признака МТ нормальному закону должна быть отвергнута и принята альтернативная гипотеза (H_1).

Аппроксимация эмпирического распределения признака МТ логнормальным распределением позволяет принять H_0 (χ^2=9,0; P=0,17), так как P>0,05.

Аппроксимация признака ИК нормальным и логнормальным распределениями показала, что этот признак не аппроксимируется ни нормальным (χ^2=19,4; P=0,0007), ни логнормальным распределениями (χ^2=10,6; P=0,03).

В соответствии с теорией обобщенных распределений эмпирическое распределение признака ДТ хорошо аппроксимируется теоретическим распределением 3-го типа (рис. 6), параметры которого указаны в табл. 2.

Параметры С (мода), А и В (точки перегиба, то есть точки, в которых вторая производная от плотности распределения равна 0) позволяют математически строго выделить центральный (от А до В) и периферические диапазоны (<A и > B) и присвоить им значения 2 балла (от А до В), 1 балл (<A) и 3 балла (>B). Такая градация диапазонов ДТ позволяет использовать данный признак для выделения соматотипов мужчин на основе полуколичественной (в баллах) оценки антропометрических данных. Судя по уровню χ^2, использование теории обобщенных распределений позволило более точно определить характер распределения признака ДТ у мужчин в возрасте от 18 до 36 лет.

Аналогичный подход был применен для аппроксимации эмпирического распределения признака МТ, ко-

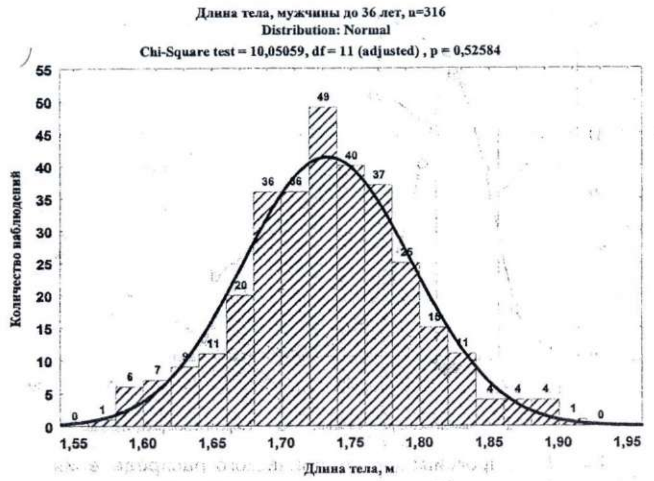

Рис. 2. Гистограмма эмпирического распределения признака ДТ и кривая ее аппроксимации нормальным распределением

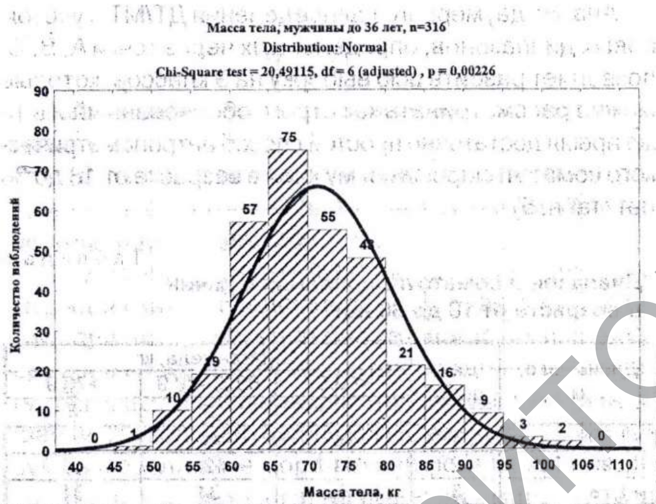

Рис. 3. Гистограмма эмпирического распределения признака МТ и кривая ее аппроксимации нормальным распределением

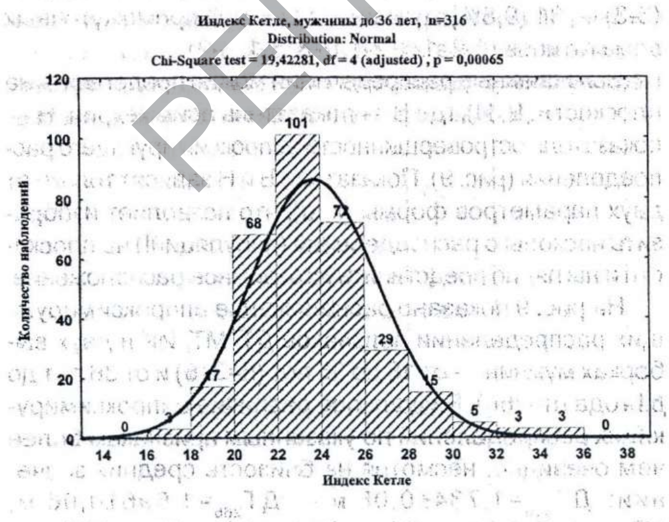

Рис. 4. Гистограмма эмпирического распределения признака ИК и кривая ее аппроксимации нормальным распределением

Таблица 1

ДТ, МТ и ИК в выборке мужчин от 18 до 36 лет (n=316)

Параметр	Возраст, лет	Длина, м	Масса, кг	ИК, кг/м^2
Mean	26,9	1,734	71,3	23,7
Median	27	1,733	69,9	23,2
SD	4,4	0,06	9,6	2,9
Min	17,0	1,580	50	17,6
Max	35,0	1,905	102	34,7
LQ_{25}	24,0	1,695	64,6	21,8
UQ_{75}	30,0	1,772	76,9	25,2

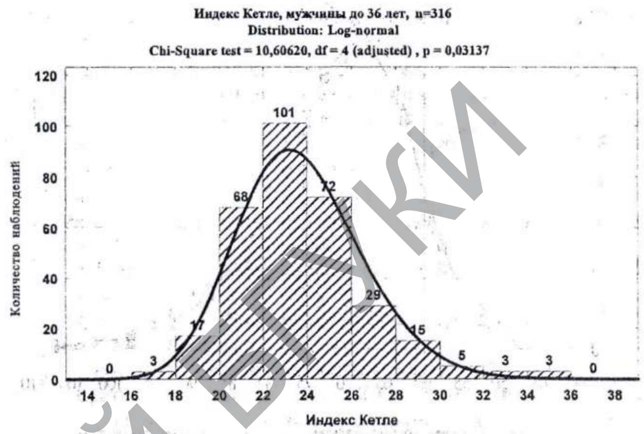

Рис. 5. Гистограмма эмпирического распределения признака ИК и кривая ее аппроксимации логнормальным распределением

Таблица 2

Параметры 3-го типа распределения признака ДТ

Параметр	Значение
αu	-1,260655754 E-14
β	57,85987
γ	44,44389
k=γ/β	0,76813
u	-1,568881
N	8,31809808 E-10
C (мода), м	1,734
A, м	1,690
B, м	1,777

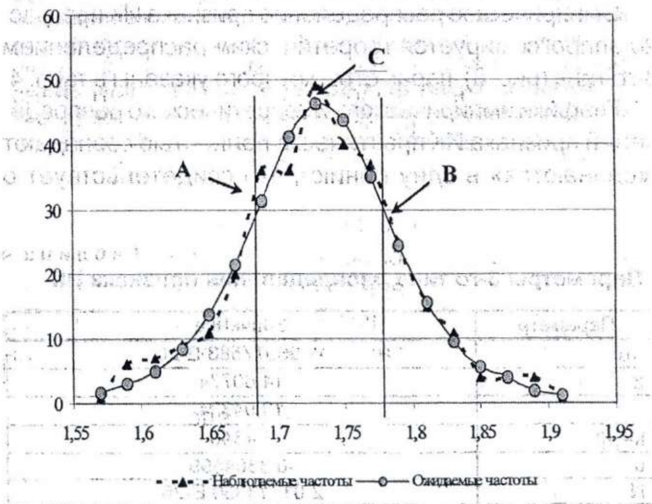

Рис. 6. Аппроксимация эмпирического распределения признака ДТ теоретическим распределением третьего типа (χ^2=4,365; P=0,9963)

торый, как было показано, нельзя аппроксимировать с помощью нормального распределения, а также признака ИК, не поддающегося выравниванию не только нормальным, но и логнормальным распределением.

В рамках теории обобщенных распределений эмпирическое распределение признака МТ хорошо аппроксимируется теоретическим распределением 3-го типа (рис. 7), параметры которого указаны в табл. 3.

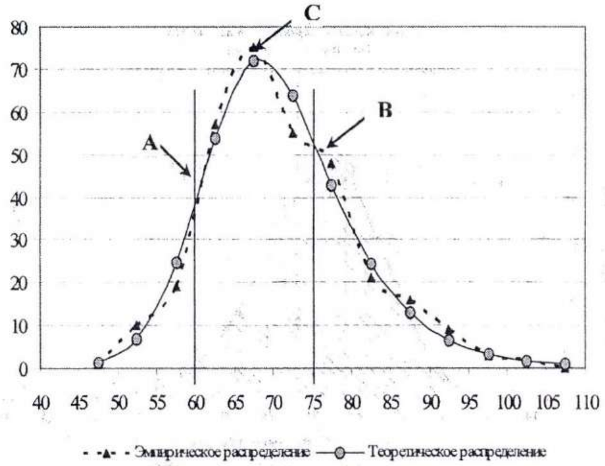

Рис. 7. Аппроксимация эмпирического распределения признака МТ теоретическим распределением 3-го типа (χ^2=6,856; Р=0,9963)

Таблица 3

Параметры 3-го типа распределения признака МТ

Параметр	Значение
αu	-1,40898714 E-16
β	8,723789
γ	21,22715
k=γ/β	2,4346
u	-0,3389107
N	1,17707527 E-37
С (мода), кг	68,3
А, кг	60,6
В, кг	75,9

Эмпирическое распределение признака ИК прекрасно аппроксимируется теоретическим распределением 3-го типа (рис. 8), параметры которого указаны в табл. 4.

Графики эмпирического и теоретического распределений признака ИК практически полностью совпадают («сливаются» в одну линию), что свидетельствует о

Таблица 4

Параметры 3-го типа распределения признака ИК

Параметр	Значение
αu	-1,96977583 E-20
β	14,60774
γ	19,94334
k=γ/β	2,4346
u	-0,8304599
N	2,01713350 E-26
С (мода), кг/м^2	22,89
А, кг/м^2	20,88
В, кг/м^2	24,87

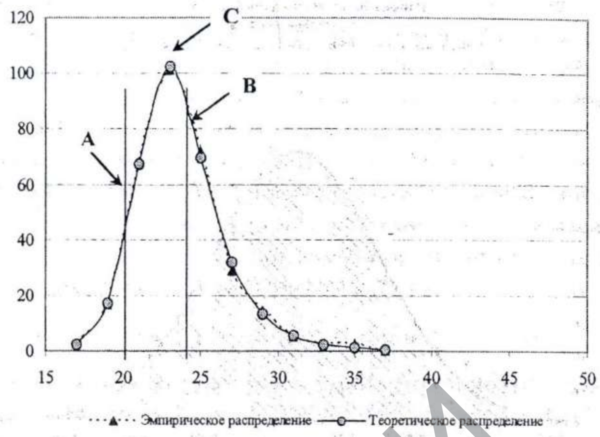

Рис. 8. Аппроксимация эмпирического распределения признака ИК теоретическим распределением третьего типа (χ^2=1,143; Р=0,9502)

хорошей аппроксимации данным типом обобщенных распределений.

Анализ двумерного распределения ДТ/МТ с учетом новых диапазонов, определенных через точки А, В, С, позволяет разбить всю выборку на 9 классов, которые можно рассматривать как строго обоснованный и в то же время достаточно простой способ антропометрического соматотипирования мужчин в возрасте от 18 до 36 лет (табл. 5).

Таблица 5

Диапазоны соматотипирования мужчин в возрасте от 18 до 36 лет

Длина тела, м	Диапазон	Масса тела, кг		
		<60,6	60,6—75,9	>75,9
		1	2	3
<1,69	1	18	40	8
1,69—1,78	2	14	121	49
>1,78	3	1	34	31

Как следует из табл. 5, предлагаемый способ соматотипирования позволяет выделить 3 основных соматотипа: лептосомный (1-1) — 18 (5,7%) мужчин, мезосомный (2-2) — 121 (39,3%) мужчина, пикносомный (3-3) — 31 (9,8%) мужчина, а также 6 промежуточных соматотипов (1-2, 1-3, 2-1, 2-3, 3-1, 3-2).

Вычисленные распределения можно представить на плоскости (В; Н), где В — показатель асимметрии, Н — показатель островершинности аппроксимирующего распределения (рис. 9). Показатели В и Н зависят только от двух параметров формы (k, u), что позволяет изобразить несколько распределений (популяций!) на плоскости и наглядно представить их взаимное расположение.

На рис. 9 показано расположение аппроксимирующих распределений признаков ДТ, МТ, ИК в двух выборках мужчин — от 18 до 36 лет (n=316) и от 36 лет до 61 года (n=266). Визуальное различие аппроксимирующих распределений по указанным признакам более чем очевидно, несмотря на близость средних значений: $ДТ_{316}$=1,734±0,06 м — $ДТ_{266}$=1,696±0,06 м; $МТ_{316}$=71,3±9,6 кг — $МТ_{266}$=72,0±11,6 кг; $ИК_{316}$=23,2±2,9 кг/м^2 — $ИК_{266}$=25,0±3,6 кг/м^2.

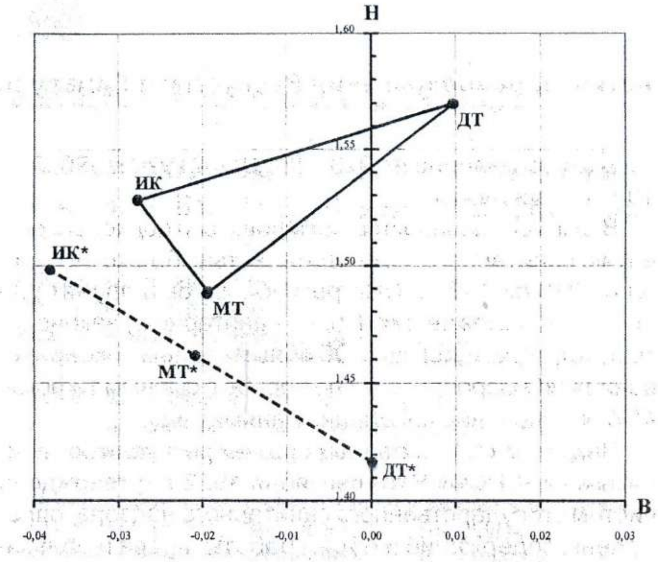

Рис. 9. Изображение выравнивающих распределений на плоскости (В; Н): ДТ, МТ, ИК — мужчины до 36 лет (n=316); ДТ*, МТ*, ИК* — мужчины старше 36 лет (n=266)

Выводы

Использование теории обобщенных распределений позволяет:

1) вычислять выравнивающее теоретическое распределение, а не «подбирать» его путем перебора известных распределений;

2) устранить извечную проблему биологических и медицинских исследований по определению «нормы» при интерпретации биомедицинских данных — среднее значение и стандартное отклонение не могут быть использованы для этих целей, так как большая часть количественных переменных, характеризующих различные морфофизиологические признаки организма человека, не имеет нормального распределения;

3) определять в соответствии с вычисленным аппроксимирующим распределением, параметры С (мода), А и В (точки перегиба — точки, в которых вторая производная от плотности распределения равна 0) и на этом основании математически строго выделять центральный (от А до В), а также периферические диапазоны (<А и > В), интерпретируя центральный диапазон как популяционную норму;

4) использовать указанные в пункте 3 диапазоны для разработки системы полуколичественного (в баллах) способа соматотипирования;

5) визуализировать аппроксимирующие распределения в виде точек на плоскости «асимметрия/эксцесс» и оценить удаленность этих точек (популяций!) друг от друга и от известных теоретических непрерывных распределений (нормальное, логнормальное, Лапласа, Коши, Вейбулла и др.).

Литература

1. Куршакова Ю. С. // Вопр. антропологии.— 1964.— Вып. 18.— С. 73—79.

2. Лакин Г. Ф. Биометрия.— М., 1990.

3. Нешитой В. В. Методы статистического анализа на базе обобщенных распределений: Учеб.-метод. пособие.— Минск, 2001.

4. Нешитой В. В. // Ученые записки Тартуского университета.— 1987.— Вып. 774.— С. 123—134.

5. Нешитой В. В. // Кибернетика.— 1987.— № 6.— С. 91—96.

6. Нешитой В. В. // Новости стандартизации и сертификации.— 2004.— № 1.— С. 54—58.

7. Нешитой В. В. Элементы теории обобщенных распределений.— Минск, 2009.

8. Орлов А. И. // Заводская лаборатория.— 1991.— Т. 57, № 7.— С. 64—66.

9. Орлов А. И. Прикладная статистика.— М., 2004.

10. Орлов А. И. // Вестн. АМН СССР.— 1987.— № 2.— С. 88—94.

11. Реброва О. Ю. Статистический анализ медицинских данных.— М., 2003.

12. Салівон І. І. Фізічны тып беларусаў: Узрастковая, тыпалагічная і экалагічная зменлівасць.— Мінск, 1994.

13. Халафян А. А. Statistica 6. Статистический анализ данных.— М., 2008.

Поступила 01.06.09.

NEW APPROACH TO STATISTICAL ANALYSIS OF BIOMEDICAL DATA

V. E. Yagur, V. V. Neshitoi, I. I. Salivon

On an example of some anthropometric characteristics (body length, body weight, Ketle index) analysis in men (n=316) aged 18—36 the possibility to assess the parameters of biomedical data empirical distributions by means of approaches developed with the theory of generalized distributions was investigated. The specified characteristics empirical distributions were well approximated by the theoretical distribution of the third type. The theoretical curve parameters — C (mode), A and B (inflection points) permitted to mark out mathematically and rigorously the central (from A to B) and the peripheral (< A and > B) ranges treating the central range as the population standard. The ranges calculated were used for developing a semi-quantitative (in points) system of somatotyping men aged 18—36. The possibility of visual rendering the approximating distributions as points on the plane "Skewness — kurtosis" permitting to estimate the remoteness of each point (population!) from the other points and from the well-known theoretical continuous distributions (normal, lognormal, Laplace, Cauchy, Weibull, etc.) had been demonstrated.

Key words: *biomedical data, anthropometric characteristics, empirical distribution, theoretical distribution, approximation curve.*