

**АВТОМАТИЧЕСКИЙ РАСЧЁТ СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ ВЫБОРКИ***Национальный исследовательский ядерный университет МИФИ*

При подготовке рукописей к публикации все редакции требуют «указать методы статистической обработки и способы представления данных» или «придерживаться Единых требований к рукописям Международного Комитета Редакторов Медицинских Журналов (ICMJE)». Но статистический анализ вызывает у врачей большие затруднения по простой и очевидной причине: «Многие вещи нам непонятны не потому, что наши понятия слабы; но потому, что сии вещи не входят в круг наших понятий» (Козьма Прутков). Поэтому разделы статей с изложением результатов анализа часто подвергаются справедливой критике [1, 2], которая, однако, вряд ли вдохновляет врачей на освоение методов анализа.

Чтобы помочь врачам самостоятельно получать достаточно полное и понятное статистическое описание выборки предлагается использовать шаблон,

созданный из функций программы Excel. Шаблон в формате *xlsx* (Excel 2010) доступен для скачивания на сайте журнала по адресу ([www.vestiar.ru](http://www.vestiar.ru)). Краткая инструкция по использованию расположена на листе с шаблоном. Копия шаблона показана в табл. 1.

Первый столбец таблицы имитирует номера строк, верхняя строка - заголовки столбцов листа книги Excel. В столбец А, начиная с ячейки А1, вводятся данные и сортируются по возрастанию (указаны 25 чисел выборки, которая обсуждается в табл. 3.) Ячейки столбца С содержат формулы обработки показателей столбца В. В строках 2-12 столбца С указаны значения основных параметров выборки. В строках 13-16 столбца В указаны дополнительные показатели выборки с числами в квадратных скобках. Если значения исследуемой выборки в столбце С не превышают значений в квадратных скобках,

выборку можно считать нормальной или близкой к нормальной. Превышение значений в квадратных скобках указывает на отклонение от нормальности [3]. В строках 17-24 столбца В указано интуитивно понятное соответствие между отклонениями от среднего и процентилями для оценки близости распределения к нормальному, предложенное в работе [4] (в литературе чаще употребляется термин перцентиль - *percentile*). В строке 25 столбца С указано рекомендуемое число интервалов гистограммы, которое полезно знать при её построении. Число знаков после запятой в столбце С устанавливается при необходимости стандартными для Excel приёмами.

Проиллюстрируем возможности использования шаблона для описания и сравнения выборок. Наиболее просто для анализа и понимания его результатов выборки с нормальным или похожим на нормальное распределениями [3, 5]. Считается, что рост людей обычно подчиняется нормальному закону распределения. Сформируем выборку здоровых подростков 18-ти лет, для которых в таблицах Всемирной организации здравоохранения (ВОЗ) указан средний рост  $M=176,14$  см и стандартное отклонения  $sd=7,47$  см в предположении, что генеральная совокупность является нормально распределённой [6].

С помощью генератора случайных чисел (Excel/Анализ данных/Генерация случайных чисел/Распределение нормальное) сформируем выборку с указанными пара-

**Таблица 1.****Расчёт параметров выборки в Excel**

	А	В	С
1	13	Показатели	Значения
2	16	Объём выборки, n	1575
3	17	Минимум	13
4	17	Максимум	96
5	19	Среднее арифметическое М	70
6	20	Стандартное отклонение sd	13,5
7	20	Доверительный интервал 95%ДИ	0,7
8	21	Медиана выборки Me	73
9	21	Перцентиль 25%	65
10	21	Перцентиль 75%	79
11	21	95% Нижняя доверительная граница Me	73
12	21	95% Верхняя доверительная граница Me	74
13	22	Коэффициент вариации $CV=sd/M$ [0,1]	0,19
14	22	Средняя квадратичная погрешность М [5%]	0,5
15	22	Асимметрия (скос) $As$ [ $\pm 0,2$ ]	-1,42
16	22	Экцесс $Ex$ [ $\pm 1$ ]	2,48
17	23	$M+sd$	84
18	23	Перцентиль 84%	81
19	24	$M-sd$	57
20	24	Перцентиль 16%	60
21	24	$M+2sd$	97
22	24	Перцентиль 97,5%	89
23	24	$M-2sd$	43
24	25	Перцентиль 2,5%	31
25	25	Число интервалов k	11,6

метрами  $M$  и  $sd$ . Для уменьшения погрешности анализа примем объём выборки  $n=100$ . Обработанные в шаблоне параметры выборки представлены в табл. 2, в которой оставлены информативные для данного примера часть строк из табл. 1.

В строках 2-3 цифры в квадратных скобках повторяют табличные значения ВОЗ. Как видно, по показателям строк 5-11 сформированную выборку можно уверенно считать нормальной (поэтому значение медианы  $Me=M$  опущено). В дополнение на рис. 1 представлена гистограмма выборки, построенная в программе Statistica [7]. Положение нормальной кривой на гистограмме и уровень значимости критерия Шапиро-Уилка  $p=0,99$  свидетельствуют о нормальности распределения. Критерий Шапиро-Уилка рекомендуется как наиболее мощный критерий нормальности распределения [5, 8].

Сравним информативность двух показателей выборки. Доверительный интервал 95%ДИ означает, что при многократном формировании выборки в 95-ти случаях из 100, т.е. с доверительной вероятностью 95% (или 0,95), среднее значение  $M=176,1$  см будет находиться в интервале  $M \pm ДИ = (174-178)$  см (с округлением). Иначе говоря, доверительные границы определяют интервал воспроизводимости результатов измерений. Про уровень значимости  $p=0,99$  критерия Шапиро-Уилка можно только сказать, что нулевая гипотеза  $H_0$  не отклоняется на уровне значимости  $p=0,99$ .

Очевидно, знание показателей одной выборки необходимо для сравнения с другой. Статистическое сравнение выборок можно осуществлять двумя методами. В первом, наиболее частом случае, по умолчанию проверяется гипотеза  $H_0$ , которая предполагает, что по изучаемому параметру выборки не различаются, т.к. они образованы случайными вариантами. Для проверки применяются определённый статистический критерий сравнения выборок и по полученному уровню значимости  $p$  делается заключение (вывод): статистическая гипотеза  $H_0$  не отклоняется, если  $p \geq 0,05$  (различия по изучаемому параметру не значимы) или отклоняется, если  $p < 0,05$  (различия по изучаемому параметру значимы) [5].

Во втором случае для изучаемого параметра вычисляется доверительный интервал на заданном уровне надёжности (обычно 95% или 0,95) и делается заключение (вывод): если доверительные интервалы  $\pm 95\%ДИ$  перекрываются, различия для изучаемого параметра статистически незначимы, если не перекрываются, различия можно считать статистически значимыми [5].

Первый метод неоднократно «...критиковали за то, что статистическая значимость наблюдаемого эффекта оценивается без учёта его величины... Эта критика привела к тому, что в настоящее время всё большее значение придаётся не самой величине  $p$ , а доверительным интервалам» [9]. Более обоснованная критика первого метода этого же автора изложена в статье с выразительным заголовком [10]. Наконец, при публикации материалов, содержащих статистический анализ, «Не следует полагаться исключительно на проверку статистических гипотез, например, на использование

$p$ -значений, которые не содержат важную информацию о размере эффекта» [11].

Рассмотрим простейший и наглядный пример сравнения выборок двумя методами, используя данные из табл. 2. Выборки для сравнения образуем из исходной, прибавляя к каждой её варианту по 1, 2, 3 см и определяя в шаблоне рис. 1 средние значения  $M$  и доверительные интервалы 95%ДИ. Результаты сравнения представлены на диаграмме рис. 2, построенной в Excel. Квадраты на диаграмме обозначают средние значения  $M$ , вертикальные линии (планки погрешностей в Excel) - доверительные интервалы  $\pm 95\%ДИ$ , пунктир - уровень верхней доверительной границы  $M+95\%ДИ$  исходной выборки.

В дополнение к ДИ оценим значимость различий между средними значениями. Все выборки в соответствии с процедурой их формирования являются нормальными и парными (зависимыми). Это позволяет определять одностороннюю ( $Difference < 0$ ) значимость различий  $p$ , используя самый мощный критерий сравнения нормальных парных выборок Equal-Variance T-Test [5, 8]. Значения  $p$  сравниваемых выборок относительно исходной определены в программе [12] и указаны внизу диаграммы.

Клиническая значимость различий в этом примере значения не имеет. Поэтому обратим внимание на коли-

Таблица 2.

Параметры выборки «Рост»

№	Показатели	Значения
1	Объём выборки, $n$	100
2	Среднее арифметическое $M$ [176,1 см]	176,1
3	Стандартное отклонение $sd$ [7,47 см]	7,88
4	Доверительный интервал 95%ДИ	1,56
5	Коэффициент вариации $CV=sd/M$ [0,1]	0,04
6	Асимметрия (скос) $As$ [ $\pm 0,2$ ]	0,03
7	Экссесс $Ex$ [ $\pm 1$ ]	-0,17
8	$M+sd$ , см	184
9	Процентиль 84% см	184
10	$M-sd$ , см	168
11	Процентиль 16%, см	168
12	Число интервалов $k$	7,6

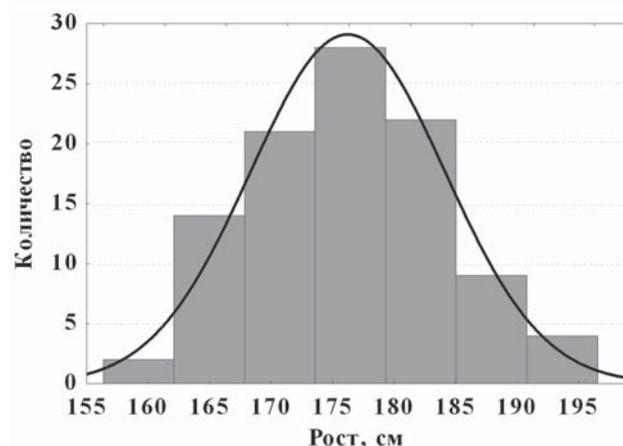


Рис. 1. Гистограмма «Рост».

чественные характеристики различий. При увеличении роста на 2 см верхняя доверительная граница исходной выборки заметно превышает нижнюю доверительную границу выборки +2 см, тогда как уровень значимости  $p=0,04$  уже указывает на наличие различий. И даже при

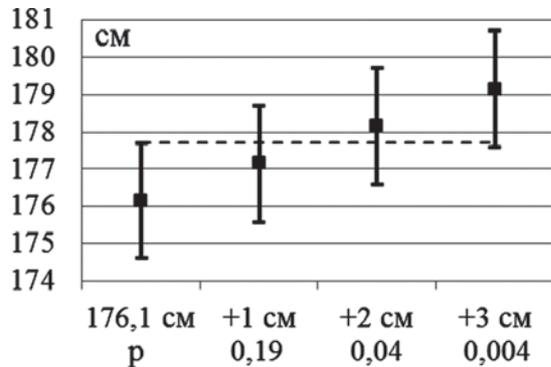


Рис. 2. Доверительные интервалы. Пояснения в тексте.

Таблица 3.

Параметры выборки «Возраст»

№	Показатель	Значения
1	Объём выборки, n	1575
2	Среднее арифметическое M	70
3	Стандартное отклонение sd	13,5
4	Доверительный интервал 95%ДИ	0,7
5	Медиана выборки Me	73
6	Процентиль 25%	65
7	Процентиль 75%	79
8	95% нижняя доверительная граница Me	73
9	95% верхняя доверительная граница Me	74
10	Коэффициент вариации $CV=sd/M [0,1]$	0,19
11	Асимметрия (скос) $As [\pm 0,2]$	-1,42
12	Экссесс $Ex [\pm 1]$	2,48
13	$M+sd$	84
14	Процентиль 84%	81
15	$M-sd$	57
16	Процентиль 16%	60
17	Число интервалов k	11,6

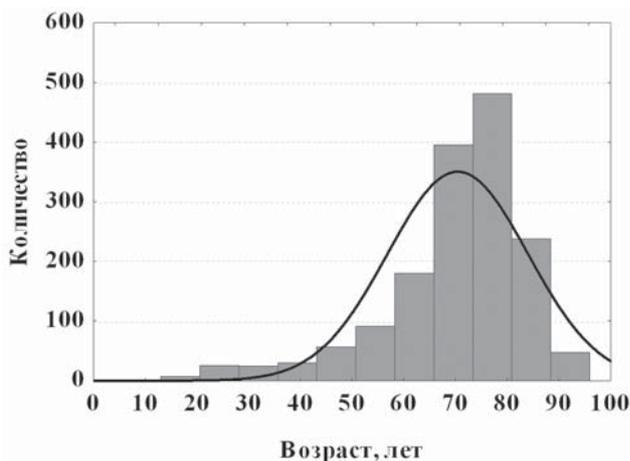


Рис. 3. Гистограмма «Возраст».

увеличении роста на 3 см доверительные границы хотя и мало, но ещё перекрываются, тогда как различия считаются высоко значимыми [5, 13]. Как видно, доверительные интервалы более наглядно отражают различия средних M на диаграмме, чем значения уровней статистической значимости p внизу диаграммы.

Отметим дополнительные преимущества статистических оценок по доверительным интервалам сравнительно с оценкой по величине p. Для доверительных интервалов можно задать доверительную вероятность. Например, для выборки из табл. 1 значение 95%ДИ = 1,56 см, а если задать доверительную вероятность 99%, то получим 99%ДИ = 2,07 см. Для критериев согласия и сравнения такое понятие отсутствует.

Доверительные интервалы при изменении объёма выборки предсказуемо и заметно изменяются. Например, если объём выборки из табл. 2 с прежними параметрами  $M \pm sd = (176,14 \pm 7,47)$  см уменьшить (генератор случайных чисел) до значения  $n=20$ , то получим 95%ДИ=3,42 см, т.е. увеличение в 2,2 раза. При этом уровень значимости критерия согласия Шапиро-Уилка составит  $p=0,77$ , т.е. уменьшение на 22%. Добавим, что при  $n \leq 20$  статистическое описание смысл утрачивает [5], а визуальная оценка выборки по гистограмме становится бесполезной.

Рассмотрим особенности расчёта доверительных интервалов выборок, распределение которых сильно отличается от нормального. Врачи часто упускают из виду, что «для большинства физиологических переменных распределения представляют собой гладкие унимодальные асимметричные кривые ... Здоровые и больные фактически принадлежат к двум разным популяциям» [14] (об этом же и многие ссылки из списка литературы). А для ненормальных выборок доверительные интервалы, рассчитанные в Excel и в статистической программе, могут различаться.

Проиллюстрируем это различие на примере реальной выборки (табл. 1) возраста пациентов, которым в Московском центре кардиостимуляции в течение одного года были имплантированы 1575 кардиостимуляторов. Информативная для данного примера часть строк из табл. 1 представлена в табл. 3.

Отметим показатели, указывающие на значительное отклонение распределения от нормального. Значение среднего M в строке 2 отличается от значения медианы Me в строке 5. В строках 10-12 существенно превышены граничные значения в квадратных скобках, что свидетельствует о ненормальности выборки [3], в строках 13-16 значения  $M \pm sd$  заметно отклоняются от соответствующих процентилей [4]. В дополнение на рис. 3 представлена гистограмма выборки. Положение нормальной кривой на гистограмме и уровень значимости критерия Шапиро-Уилка  $p=0,0000$  свидетельствуют о ненормальности распределения.

В большинстве публикаций клинических результатов со статистической оценкой, прежде всего, указывается возраст пациентов, обычно в формате  $M \pm sd$  и без проверки нормальности распределения. Но для данной выборки стандартное отклонение sd не является мерой её рассеяния из-за расхождения значений M и Me, но ещё больше из-за расхождения  $M \pm sd$  с

перцентильями 84% - 16% и  $M \pm 2sd$  с перцентильями 97,5% - 2,5% (табл. 1 с этой же выборкой). При значительных асимметрии и эксцессе в качестве меры центральной тенденции рекомендуется использовать медиану  $Me$ , которая более устойчива к аномальным выбросам, а в качестве меры рассеяния перцентили (обычно интерквартильный размах). В этой выборке возраст пациентов составляет  $Me$  (25%; 75%) = 73 (65; 79) лет (перцентили 25%; 75% указаны в табл. 1). По этим же причинам значение доверительного интервала из строки 4 табл. 3 уже не характеризует воспроизводимость медианы. Добавим, что в ячейках С9, С10 табл. 1 легко задать значения других перцентилей, например  $Me$  (5%; 95%) = 73 (42; 87) лет.

Воспроизводимость медианы удобнее характеризовать нижней и верхней доверительными границами (ДГ) из-за возможной асимметрии этих границ отно-

сительно самой медианы. В строках 8-9 (11-12 в табл. 1) доверительные границы с доверительной вероятностью 95% рассчитаны по формулам из книги [8]. В этой выборке значение медианы с ДГ составляет  $Me=73$  (73-74) лет. ДГ медианы позволяют рассчитывать некоторые статистические программы, например [12].

Для выборок объемом  $n \leq 100$  номера строк (рангов) нижней и верхней ДГ, в которых находятся сами ДГ, можно определить по таблицам в книге [5] или в ГОСТ'e [15]. Выборку необходимо предварительно отсортировать по возрастанию.

При сравнении выборок одна из них может иметь распределение нормальное, а другая - ненормальное. В этом случае показатели меры центральной тенденции и меры рассеяния у обеих выборок должны быть, очевидно, одинаковыми. Какие именно выбрать показатели решает врач.

### ЛИТЕРАТУРА

1. Леонов В.П. Три «Почему ...» и пять принципов описания статистики в биомедицинских публикациях. 2006. Биометрика-Томск. // Сайт биостатистики в медицине и биологии. <http://www.biometrica.tomsk.ru/principals.htm>
2. Гржибовский А.М. Использование статистики в российской биомедицинской литературе. // Экология человека, 2008, № 12, с.55-64.
3. Львовский Е.Н. Статистические методы построения эмпирических формул. // М., «Высшая школа», 1988, 239 с.
4. Гланц С. Медико-биологическая статистика. Пер. с англ. // М., Практика, 1998, 459 с.
5. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. // М., МедиаСфера, 2002. 312с.
6. Сайт ВОЗ [http://www.who.int/growthref/hfa\\_boys\\_5\\_19years\\_z.pdf?ua=1](http://www.who.int/growthref/hfa_boys_5_19years_z.pdf?ua=1)
7. STATISTICA v. 8.0, StatSoft, Inc.
8. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. // М., Физматлит, 2006, 816 с.
9. Гудман С.Н. На пути к доказательной биостатистике. Часть 1: обманчивость величины  $p$ . // Международный журнал медицинской практики, 2002, № 1, с. 8-17 (Translated from Ann. Intern. Med. 1999) <http://www.mediasphera.ru/mjamp/2002/1/r1-02con.htm>
10. Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. // Semin. Hematol., 2008, Jul,45(3), p.135-40. <http://xa.yimg.com/kq/groups/18751725/636586767/name/twelve+P+value+misconceptions.pdf>.
11. Международный комитет редакторов медицинских журналов: Единые требования к рукописям, представляемым в биомедицинские журналы: правила написания и редактирования материалов. // Межд. журн. мед. практики, 2005, № 5, с. 10-23. <http://www.mediasphera.ru/mjamp/2005/5/10.pdf>.
12. NCSS 2004 and PASS 2005. Statistical Software.
13. StatSoft. Электронный учебник по статистике. Элементарные понятия статистики. [http://www.statsoft.ru/home/textbook/esc.html#What is «statistical significance» \(p-level\)](http://www.statsoft.ru/home/textbook/esc.html#What%20is%20statistical%20significance%20(p-level))
14. Флетчер Р., Флетчер С., Вагнер Э. Клиническая эпидемиология. Основы доказательной медицины. // М., Медиа Сфера, 1998, 352 с.
15. ГОСТ Р ИСО 16269-7-2004. Статистическое представление данных. Медиана. Определение точечной оценки и доверительных интервалов. // М., ИПК Издательство стандартов, 2004, 11 с. <http://docs.cntd.ru/document/1200035332>