УДК 519.25

ОПИСАТЕЛЬНАЯ СТАТИСТИКА И ПРОВЕРКА НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЯ КОЛИЧЕСТВЕННЫХ ДАННЫХ

© 2014 г. [1,2]А. В. Субботина, [1-3]А. М. Гржибовский

[1]Университет г. Тромсё, Норвегия
[2]Архангельская международная школа общественного здоровья ГБУО ВПО «Северный государственный медицинский университет», г. Архангельск, Россия
[3]Норвежский институт общественного здравоохранения, г. Осло, Норвегия

В статье представлены основные типы данных, принципы проверки их распределения с использованием пакета прикладных статистических программ Stata, а также описательная статистика для количественных переменных. На примере реальных данных рассматриваются основные принципы их первичной обработки и приемы описательной статистики, основной задачей которой является описание полученных в ходе исследования данных в максимально сжатом виде с минимальной потерей информации. Работа продолжает начатую в первом номере журнала серию статей по основам биостатистики с применением пакета программ для статистической обработки данных Stata. Однотипное изложение материала позволит читателям, знакомым с нашими предыдущими публикациями, на аналогичных примерах изучить использование новой для них программы. Настоящая серия статей может быть также использована как самостоятельное пособие по статистической обработке данных биомедицинских исследований.

Ключевые слова: типы данных, нормальное распределение, описательная статистика

Статья открывает серию публикаций, посвященных статистической обработке данных биомедицинских исследований в статистической программе Stata 12.0. Однотипное с предыдущими публикациями в «Экологии человека» на тему использования пакета статистических программ SPSS изложение материала позволит читателям, знакомым с указанной серией статей, на аналогичных примерах изучить использование новой для них программы Stata, которая обладает большим спектров возможностей, чем SPSS. Поскольку ранее опубликованные пособия по основам биостатистики [1–9] не содержат информации о применении программы Stata, наши статьи могут быть использованы как самостоятельное руководство по статистической обработке данных биомедицинских исследований. Однако следует отметить, что в них будет сделан акцент на прикладные аспекты исследований – основное внимание будет уделено правилам выбора наиболее подходящего способа обработки данных в зависимости от имеющегося материала, алгоритмам применения выбранных статистических процедур статистической программы Stata и интерпретации полученных результатов. Предполагается, что читатель знаком с основными моментами теории медицинской статистики и эпидемиологических исследований. Для более глубокого изучения рассмотренных вопросов читателю предлагается обращаться к специальной литературе по биомедицинской статистике.

Переменной называют величину, характеризующуюся множеством значений, которое она может принимать. Переменные (изучаемые признаки, variables) делятся на количественные и качественные. Количественными являются переменные, количественная мера которых четко определена, качественными – переменные, не поддающиеся числовому измерению.

Количественные переменные подразделяются на непрерывные (continuous) и дискретные (discrete). Непрерывные переменные могут принимать любое значение на непрерывной шкале, например, рост, масса тела, артериальное давление, биохимические показатели крови и т. д. Дискретные переменные могут выражаться только целыми числами, например, количество детей в семье, число выкуриваемых сигарет в день, количество рецидивов заболевания и т. д.

Качественные переменные, в свою очередь, делятся на номинальные (nominal, categorical) и порядковые или ранговые (ordinal). К номинальным переменным относятся характеристики, с которыми нельзя производить арифметические действия и которые нельзя расположить в порядке возрастания или убывания, например, идентификационный номер пациента, диагноз, название препарата, семейное положение и т. д. Порядковые (ранговые) переменные можно расположить (ранжировать) в логическом порядке, например, стадия болезни или оценка тяжести

состояния пациента, однако невозможно количественно выразить, насколько или во сколько раз одно состояние лучше или хуже другого. Несмотря на то, что при занесении порядковых переменных в компьютер их часто кодируют с помощью цифр, с ними, в отличие от количественных данных, нельзя производить арифметические действия. Например, оценка на экзамене в университете (по пятибалльной шкале) является типичным примером порядковой величины. Мы знаем, что оценка «отлично», традиционно выражаемая в виде «5», лучше, чем оценка «хорошо», выражаемая в виде «4», а оценка «удовлетворительно» или «3» лучше, чем оценка «неудовлетворительно» или «2», однако мы не можем сказать, что «5» лучше, чем «4», настолько же, насколько «4» лучше, чем «3», или «3» лучше, чем «2».

Переменные, которые могут быть отнесены к противоположным категориям, то есть могут принимать только одно из двух значений (здоров/болен, умер/выжил, курит/не курит и т. д.), называются дихотомическими (dichotomous) или бинарными (binary).

Количественные данные при необходимости могут быть представлены в виде ранговых или номинальных. Например, индекс массы тела измеряется на непрерывной шкале, однако можно провести разделение выборки на лиц с недостаточной, нормальной и избыточной массой тела, создав, таким образом, порядковый признак. В дальнейшем признак можно превратить в дихотомический, объединив, к примеру, первую и вторую группы.

Ранговые переменные можно представить в виде номинальных, но не наоборот. В некоторых случаях, например при применении визуально-аналоговых шкал, ранговые переменные представляют и анализируют как количественные, однако в таких случаях следует с большой осторожностью относиться к интерпретации результатов, так как различия между значениями на одном конце шкалы (например, между 1 и 2) могут быть более выражены, чем на другом (например, между 9 и 10), несмотря на то, что числовое значение различий в обоих случаях равно единице.

Перед тем как описывать количественные данные, всегда следует проводить проверку распределения. Под видом распределения понимают функцию, связывающую значения переменной случайной величины с вероятностью их появления в совокупности [6]. В биомедицинских исследованиях чаще всего проводится «проверка распределения на нормальность». Под нормальным распределением понимают симметричное распределение колоколообразной формы, при котором около 68 % данных отличается от среднего арифметического не более чем на одно, а примерно 95 % – не более чем на два стандартных отклонения в каждую сторону. Несмотря на то, что нормальное (Гауссово) распределение встречается очень часто и играет важную роль в статистике, существуют и другие распределения данных (биномиальное, Пуассона, Максвелла, Шарлье и др.), о которых можно прочитать в специальной литературе. Проверка распределения производится тремя способами: с помощью описательной статистики, графически и с использованием статистических критериев.

Все способы проверки рассматриваются на намеренно измененном материале, полученном в ходе Северодвинского когортного исследования [11]. Файл с данными (Human_Ecology_2014_2.dta) доступен на сайте журнала. Проверим распределение семейного дохода (переменная dohod) в семьях первородящих женщин г. Северодвинска Архангельской области и массы тела их новорожденных детей (переменная ves). В файле содержатся данные только по детям, рожденным в срок 37–42 недели.

После запуска программы Stata откроем файл Human_Ecology_2014_2.dta (расширение, используемое для баз данных Stata).

Рабочее пространство программы содержит следующие окна:

Review, которое содержит список использованных команд; с помощью щелчка мышью возможен повторный вызов использованной ранее команды;

Results, основное окно, где отображаются результаты заданных команд;

внизу экрана по умолчанию находится окно Command, или командная строка;

окна Variables, содержащее названия переменных и их описание; и Properties, в котором приводятся более подробные сведения о переменных.

Меню раскрывающихся окон располагается вверху экрана.

Статистическую программу Stata отличает удобство использования командной строки. В командной строке печатаются команды, синтаксис которых состоит из названия команды и списка переменных, которые необходимо включить в анализ. Также действия доступны из раскрывающихся меню, напоминающих подобные в программе SPSS. Использование командной строки повышает эффективность использования возможностей программы. В данной серии статей мы будем рассматривать управление в основном с помощью командной строки, однако для пользователей, предпочитающих раскрывающиеся меню, будут предложены способы их использования.

Одним из неоспоримых преимуществ Stata является наличие встроенной справки, в которой, кроме детального описания синтаксиса всех имеющихся команд, приводятся примеры их использования и сопутствующие комментарии. С помощью команды

- `help`

открывается список основных разделов справки с возможностью перехода в интересующие разделы с помощью гиперссылок. Для доступа к описанию отдельных команд используется тот же синтаксис, но с указанием интересующей нас команды, например:

- `help help`
- `help findit`

Если мы не знаем точного названия команды, поиск по отдельному слову или сочетанию слов можно осуществить с помощью взаимозаменяемых команд

(слово «word» должно быть заменено на интересующую команду)

- `findit word`
- `search word`

Все команды Stata имеют аналогичный синтаксис, то есть «язык», с помощью которого задаются действия. Названия команд, как правило, совпадают с названием статистического теста или действия. С накоплением опыта использования программы необходимость использования файла справки становится все меньше, однако на этапе освоения программы рекомендуется почаще сверяться с ним. При ошибочном написании команды в окне Results появляется отмеченная красным строка с указанием причины ошибки.

Для получения описательной статистики в Stata существует набор команд. В случае, когда мы не указываем конкретной переменной или нескольких переменных, программа выводит статистику для всего набора переменных.

Команда codebook выводит в окно результатов описание переменных (табл. 1). Она представляет такие параметры, как тип переменной, разброс, количество уникальных значений, наличие пропущенных данных, а также основные показатели описательной статистики, то есть среднее значение переменной, стандартное отклонение и значения перцентилей. С помощью раскрывающихся меню можно выполнить команду следующим образом: Data > Describe data > Describe data contents (codebook).

Это же действие выполняется при написании в командной строке команды:

- `codebook`

Таблица 1

Описательная статистика для набора переменных

dohod **(unlabeled)**

type:	numeric (**long**)		
range:	**[1900,59000]**	units:	**100**
unique values:	**159**	missing .:	**0/869**
mean:	**11966.6**		
std. dev:	**7161.34**		

percentiles:	10%	25%	50%	75%	90%
	5300	**7500**	**10200**	**14300**	**20000**

ves **(unlabeled)**

type:	numeric (**int**)		
range:	**[1900,4720]**	units:	**1**
unique values:	**332**	missing .:	**0/869**
mean:	**3388.2**		
std. dev:	**435.806**		

percentiles:	10%	25%	50%	75%	90%
	2840	**3100**	**3370**	**3660**	**3960**

Во многом дублирует предыдущую команду следующая (табл. 2):

- `summarize`

Таблица 2

Суммарная статистика

Variable	Obs	Mean	Std. Dev.	Min	Max
ves	**869**	**3388.196**	**435.8062**	**1900**	**4720**
dohod	**869**	**11966.63**	**7161.336**	**1900**	**59000**

При помощи раскрывающегося списка меню данную команду можно вызвать следующим способом: Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics

Для получения более детальной описательной статистики необходимо через запятую указать опцию detail (табл. 3 и 4).

- `summarize dohod, detail`

Таблица 3

Описательная статистика для переменной «dohod»

dohod

	Percentiles	Smallest		
1%	**2700**	**1900**		
5%	**4300**	**1900**		
10%	**5300**	**2000**	Obs	**869**
25%	**7500**	**2000**	Sum of Wgt.	**869**
50%	**10200**		Mean	**11966.63**
		Largest	Std. Dev.	**7161.336**
75%	**14300**	**49000**		
90%	**20000**	**51000**	Variance	**5.13e+07**
95%	**25000**	**53000**	Skewness	**2.191738**
99%	**40000**	**59000**	Kurtosis	**10.42976**

- `summarize ves, detail`

Таблица 4

Описательная статистика для переменной «ves»

ves

	Percentiles	Smallest		
1%	**2325**	**1900**		
5%	**2730**	**2140**		
10%	**2840**	**2200**	Obs	**869**
25%	**3100**	**2200**	Sum of Wgt.	**869**
50%	**3370**		Mean	**3388.196**
		Largest	Std. Dev.	**435.8062**
75%	**3660**	**4618**		
90%	**3960**	**4630**	Variance	**189927**
95%	**4130**	**4640**	Skewness	**.1003874**
99%	**4460**	**4720**	Kurtosis	**3.137153**

В таблицах представлены значения медианы (50 %) и квантилей, минимальное (Minimum) и максимальное (Maximum) значения переменной, значения средней арифметической (Mean), стандартного отклонения (Std. Dev.), дисперсии (Variance), а также коэффициенты асимметрии (Skewness) и эксцесса (Kurtosis).

При нормальном распределении, которое симметрично, значения медианы и среднего арифметического будут одинаковы, а значения асимметрии и эксцесса равны нулю. Если средняя арифметическая больше медианы, а коэффициент асимметрии > 0, то распределение имеет правостороннюю асимметрию (скошено вправо). При левосторонней асимметрии средняя арифметическая меньше медианы, а коэффициент асимметрии < 0. По величине коэффициента эксцесса говорят об островершинном (Kurtosis > 0) или плосковершинном (Kurtosis < 0) распределении. Однако ситуаций, когда средняя арифметическая равна медиане, а коэффициенты асимметрии и эксцесса равны нулю, практически не встречается, поэтому необходимо решить, какие отклонения от идеаль-

ного сценария допустимы для того, чтобы считать распределение полученных данных нормальным или близким к нормальному. Stata позволяет провести также формальный тест на нормальное распределение на основе вышеупомянутых показателей Skewness/Kurtosis (табл. 5 и 6).

Statistics > Summaries, tables, and tests > Distributional plots and tests > Skewness and kurtosis normality test

- `sktest dohod`

Таблица 5

Результат проверки распределения переменной «dohod» с помощью статистического критерия Skewness/Kurtosis

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
dohod	**869**	**0.0000**	**0.0000**	**.**	**0.0000**

- `sktest ves`

Таблица 6

Результат проверки распределения переменной «dohod» с помощью статистического критерия Skewness/Kurtosis

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
ves	**869**	**0.2242**	**0.3653**	**2.30**	**0.3172**

Данный тест используется при любых объемах выборки. Значимый тест ($p > 0,05$) указывает на значимое отличие распределения переменной от нормального (значение p в данном случае обозначается как Prob>chi2). Соответственно при $p < 0,05$ мы принимаем нулевую гипотезу о том, что распределение переменной не отличается значимо от нормального распределения.

Другим статистическим критерием, используемым в Stata для проверки нормальности распределения данных, является тест на проверку распределения с помощью критериев Shapiro-Wilk. Данный тест предпочтителен для использования при небольших объемах выборки, а с увеличением количества наблюдений достоверность его снижается. При объеме выборки более 2 000 наблюдений тест не достигает уровня статистической значимости. Данный тест также доступен с помощью меню (Statistics > Summaries, tables, and tests > Distributional plots and tests > Shapiro-Wilk normality test) или командной строки.

- `swilk dohod`
- `swilk ves`

При применении критериев Shapiro-Wilk за нулевую гипотезу принимается гипотеза о том, что изучаемое распределение не отличается от нормального, значит, если достигнутый уровень значимости при проверке гипотезы будет меньше, чем критический уровень значимости (p, обычно 0,05), обозначаемый в данном случае как Prob>z, то нулевая гипотеза о сходстве распределений отклоняется, значит, распределение отличается от нормального. Соответственно если $p > 0,05$, распределение не отличается от нормального. Результаты проверки гипотез о соответствии распределения переменных «dohod» и «ves» нормальному представлены в табл. 7 и 8.

Таблица 7

Результат проверки распределения переменной «dohod» с помощью статистического критерия Shapiro-Wilk

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
dohod	**869**	**0.82523**	**96.998**	**11.263**	**0.00000**

Таблица 8

Результат проверки распределения переменной «ves» с помощью статистического критерия Shapiro-Wilk

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
ves	**869**	**0.99754**	**1.365**	**0.767**	**0.22162**

Согласно значениям обоих формальных тестов, достигнутый уровень значимости для переменной «dohod» представляет собой малую величину ($p < 0,001$) и позволяет отвергнуть нулевую гипотезу о подчинении данных закону нормального распределения. Для переменной «ves» нулевую гипотезу при критическом уровне значимости 0,05 отвергнуть нельзя, значит, можно сделать вывод о том, что масса тела новорожденных в исследуемой выборке подчиняется закону нормального распределения.

К аналогичному заключению можно прийти на основании результатов анализа графиков. Гистограммы обеих переменных представлены на рис. 1 и 2.

- `histogram dohod, normal`

(Graphics > Histogram)

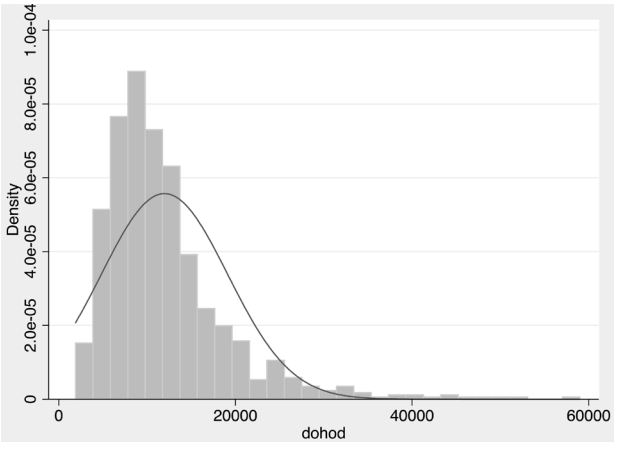

Рис. 1. Гистограмма переменной «dohod»

Гистограмма показывает, что распределение переменной «dohod» смещено вправо, что соответствует результатам описательной статистики. Непрерывная линия на рисунке показывает нормальное распределение при значениях средней арифметической и стандартного отклонения, полученных для имеющихся данных. Таким образом, гистограмма наглядно показывает, что распределение доходов в семьях не

подчиняется закону нормального распределения, а значение средней арифметической больше медианы из-за более высоких доходов небольшого количества семей. Нелишне упомянуть, что в масштабах страны распределение доходов еще сильнее смещено вправо, чем на рис. 1, то есть средние значения доходов не являются реальным отражением доходов большинства населения.

- `histogram ves, normal`

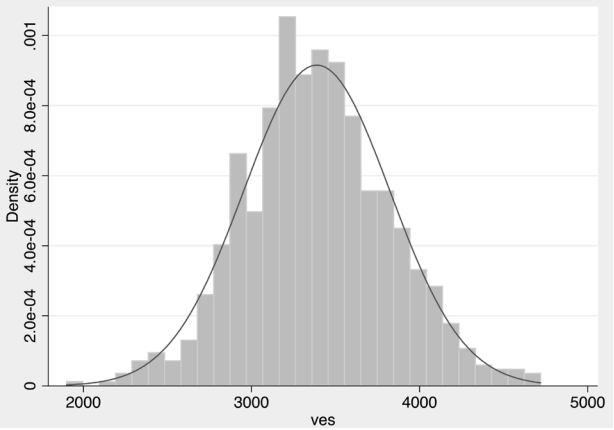

Рис. 2. Гистограмма переменной «ves»

Гистограмма для переменной «ves» имеет симметричный вид вокруг средней величины, и большинство частот находится под кривой нормального распределения, что было ранее показано с помощью критерия Shapiro-Wilk.

Несмотря на то, что гистограмма является хорошим способом проверки нормальности распределения, автоматическое создание программой шкалы может привести к неверным выводам. Более четкую картину распределения данных и соответствия распределения данных закону нормального распределения дают квантильные диаграммы (Q-Q plots).

В случае нормального распределения квантильная диаграмма имеет вид прямой линии. Любое отклонение от прямой свидетельствует об отклонении данных от нормальности. Распределение переменной «dohod» значительно отличается от нормального, что подтверждается квантильной диаграммой, на которой прямой линией обозначено, как выглядело бы нормальное распределение, а фактическое распределение видимо отклоняется от этой прямой (рис. 3).

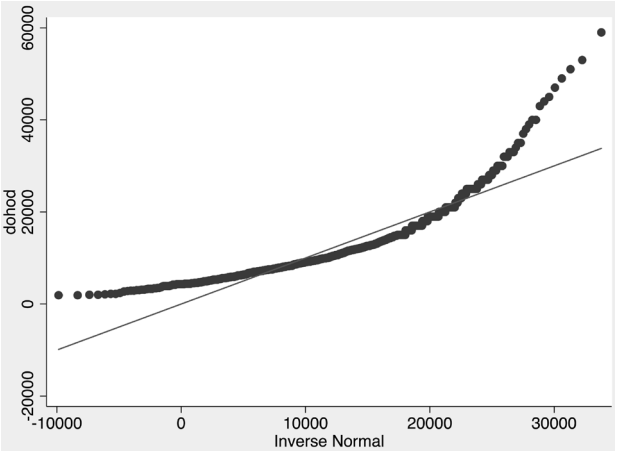

Рис. 3. Квантильная диаграмма переменной «dohod»

Statistics > Summaries, tables, and tests > Distributional plots and tests > Quantile-quantile plot

- `qnorm dohod`

Для переменной «ves» большинство значений переменной находится на прямой линии, что говорит о близости фактического распределения к нормальному (рис. 4), что было ранее показано с помощью критерия Shapiro-Wilk.

- `qnorm ves`

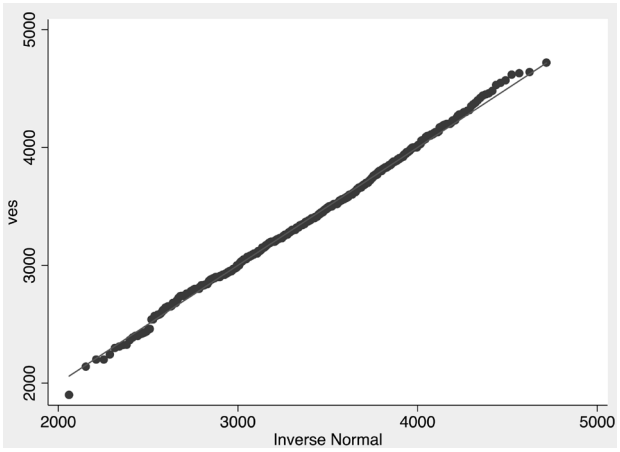

Рис. 4. Квантильная диаграмма переменной «ves»

Итак, на основании проверки распределения можно сделать вывод, что значения переменной «dohod» не подчиняются закону нормального распределения. Такие данные не рекомендуется описывать с помощью средней арифметической и стандартного отклонения, которые сильно подвержены влиянию крайних вариант. Медиана значительно меньше подвержена такому воздействию, а потому рекомендуется для описания асимметричных распределений. В качестве мер рассеяния рекомендуется использовать процентили (25-й и 75-й, называемые также нижним и верхним квартилями, соответственно используются чаще других), а также размах вариации. Таким образом, переменную «dohod» можно описать следующим образом: уровень семейных доходов в выборке варьировал от 1 900 до 59 000 рублей в месяц (*Me* = 10 200), причем доходы 50 % семей находились в промежутке от 7 500 до 14 300 рублей в месяц. Верхний и нижний квартили также представляют в виде Q1 и Q3 соответственно.

Распределение значений переменной «ves» в выборке подчинялось закону нормального распределения, а потому может описываться с помощью средней арифметической и стандартного отклонения: *M* = 3 388 г, *SD* = 436 г. Интервальную оценку популяционной средней можно представить с помощью доверительных интервалов (ДИ). В данном

примере вес детей, рожденных в срок 37–42 недели у первородящих женщин г. Северодвинска, составляет 3 388 (95 % ДИ: 3 359–3 417) г. Многие авторы представляют выборочные данные в виде $M \pm m$, где M – средняя арифметическая, а m – стандартная ошибка средней величины. Желание представлять среднюю ошибку средней арифметической вместо стандартного отклонения понятно, так как она в $\sqrt{n}$ раз меньше последнего и может маскировать существенный разброс данных вокруг среднего значения, особенно при асимметричных распределениях. Поэтому рекомендуется всегда сначала проверять распределение данных и в случае нормального распределения представлять выборочные данные в виде средней арифметической и стандартного отклонения. Асимметричные распределения лучше описывать с помощью медианы, процентилей и размаха вариации. Помимо медианы можно для описания центральных тенденций использовать моду (Mo). Мода представляет собой наиболее часто встречающееся значение переменной. Помимо самой моды рекомендуется представлять, в какой доле случаев переменная принимает значение, равное моде.

Одним из способов оценки вариабельности признака является расчет коэффициента вариации (coefficient of variation, CV), который легко получить путем деления стандартного отклонения на среднюю арифметическую с последующим умножением результата на 100 %.

```
• summarize dohod
• di 100 * r(sd) / r(mean)
```

Функция r(name) возвращает значения, рассчитанные программой в предыдущем действии. Аналогично мы могли бы подставить значения стандартного отклонения и средней арифметической в формулу

```
• di 100 * 7161.336 / 11966.63
• summarize ves
• di 100 * r(sd) / r(mean)
```

Данные считаются достаточно однородными при CV < 10 % [7], однако это разделение достаточно условно. Коэффициент вариации может применяться для сравнения разброса данных, имеющих разные размерности. Для семейного дохода CV = 60 %, в то время как для веса новорожденных CV = 13 %, то есть можно говорить о том, что разброс доходов в изучаемой выборке варьирует в несколько раз сильнее, чем разброс веса новорожденных.

Результаты проверки распределения с помощью статистических критериев всегда следует интерпретировать с осторожностью, так как они чувствительны к объемам выборок. Вероятность получения статистически значимых различий при проверке распределения при одинаковом отклонении фактического распределения от нормального при $n = 1\,000$ значительно выше, чем, скажем, при $n = 30$. Некоторые исследователи [10] рекомендуют всегда считать распределение отличающимся от нормального при $n < 30$. При условии $30 < n < 100$, если статистически критерии покажут отклонение распределения от нормального ($p < 0{,}05$), следует считать, что распределение отличается от нормального, если графики и значения асимметрии и эксцесса не свидетельствуют об обратном. При условии $n \geq 100$, если нулевую гипотезу о соответствии распределения нормальному отклонить нельзя ($p > 0{,}05$), распределение считают нормальным, если графики и значения асимметрии и эксцесса не говорят о противоположном. Для условного соответствия распределения нормальному допускается нахождение показателей асимметрии и эксцесса в интервале от –1 до 1 [10], хотя встречается и более консервативный подход, согласно которому допускаются значения асимметрии и эксцесса от –0,5 до 0,5 [3]. Авторы придерживаются мнения, что всегда следует проверять распределение несколькими способами, из которых оценка квантильной диаграммы представляется наиболее информативным.

В следующем выпуске будет рассматриваться сравнение данных, имеющих нормальное распределение в двух независимых группах.

Список литературы

1. *Банержи А.* Медицинская статистика понятным языком: вводный курс. М. : Практическая медицина, 2007. 287 с.

2. *Власов В. В.* Эпидемиология : учебное пособие для вузов. М. : ГЭОТАР-МЕД, 2004. 464 с.

3. *Жижин К. С.* Медицинская статистика : учебное пособие. Ростов н/Д : Феникс, 2007. 160 с.

4. *Наследов А. Д.* SPSS: Компьютерный анализ данных в психологии и социальных науках. СПб. : Питер, 2007. 416 с.

5. *Петри А., Сэбин К.* Наглядная статистика в медицине. М. : ГЭОТАР-МЕД, 2003. 144 с.

6. *Сергиенко В. И., Бондарева И. Б.* Математическая статистика в клинических исследованиях. М. : ГЭОТАР-МЕД, 2001. 256 с.

7. *Сырцова Л. Е., Косаговская И. И., Авксентьева М. М.* Основы эпидемиологии и статистического анализа в общественном здоровье и управлении здравоохранением : учебное пособие для ординаторов и аспирантов. М. : ММА им. И. М. Сеченова, 2003. 91 с.

8. *Таганов Д.* SPSS: Статистический анализ в маркетинговых исследованиях. СПб. : Питер, 2005. 192 с.

9. *Флетчер Р., Флетчер С., Вагнер Э.* Клиническая эпидемиология: Основы доказательной медицины. М. : МедиаСфера, 1998. 345 с.

10. *Chang Y. H.* Biostatistics 101: Data presentation // Singapore Medical Journal. 2003. N 6. P. 280–285.

11. *Grjibovski A. M., Bygren L. O., Svartbo B., Magnus P.* Social variations in fetal growth in Northwest Russia: an analysis of medical records // Annals of Epidemiology. 2003. N 9. P. 599–605.

References

1. Banerjee A. *Meditsinskaya statistika ponyatnym yazykom: vvodnyi kurs* [Medical Statistics Made Clear]. Moscow, 2007, 287 p.

2. Vlasov V. V. *Epidemiologiya* [Epidemiology]. Moscow, 2004, 464 p.

3. Zhizhin K. S. *Meditsinskaya statistika* [Medical Statistics]. Rostov-on-Don, 2007, 160 p.

4. Nasledov A. D. *SPSS: Komp'yuternyi analiz dannykh v psikhologii i sotsial'nykh naukakh* [SPSS Computer Data Analysis in Psychology and Social Sciences]. Saint Petersburg, 2007, 416 p.

5. Petrie A., Sabin K. *Naglyadnaya statistika v meditsine* [Medical Statistics at Glance]. Moscow, 2003, 144 p.

6. Sergienko V. I., Bondareva I. B. *Matematicheskaya statistika v klinicheskikh issledovaniyakh* [Mathematical statistics in clinical research]. Moscow, 2001, 256 p.

7. Syrtsova L. E., Kosagovskaya I. I., Avksent'eva M. M. *Osnovy epidemiologii i statisticheskogo analiza v obshchestvennom zdorov'e i upravlenii zdravookhraneniem* [Basics of epidemiology and statistical analysis in public health and health]. Moscow, 2003, 91 p.

8. Taganov D. *SPSS: Statisticheskii analiz v marketingovykh issledovaniyakh* [SPSS: Statistical analysis in marketing research]. Saint Petersburg, 2005, 192 p.

9. Fletcher R., Fletcher S., Vagner E. *Klinicheskaya epidemiologiya: Osnovy dokazatel'noi meditsiny* [Clinical Epidemiology: the Essentials]. Moscow, 1998, 345 p.

10. Chang Y. H. Biostatistics 101: Data presentation. *Singapore Medical Journal*. 2003, 6, pp. 280-285.

11. Grjibovski A. M., Bygren L. O., Svartbo B., Magnus P. Social variations in fetal growth in Northwest Russia: an analysis of medical records. *Annals of Epidemiology*. 2003, 9, pp. 599-605.

DESCRIPTIVE STATISTICS AND NORMALITY TESTING FOR QUANTITATIVE DATA

[1,2]A. V. Subbotina, [1-3]A. M. Grjibovski

[1]University of Troms , Troms , Norway
[2]International School of Public Health, Northern State Medical University, Arkhangelsk, Russia
[3]Norwegian Institute of Public Health, Oslo, Norway

In this article we explain types of data, main principles of testing for normality and descriptive statistics for quantitative data using Stata software. Using real examples, we present step-by-step analysis of data and presentation of descriptive statistics. This article continues a series of publications on basic biostatistics initiated in the previous issue of the journal. This series of papers can be used as a manual for analyzing the data from biomedical studies.

Keywords: types of data, normal distribution, descriptive statistics

Контактная информация:

Гржибовский Андрей Мечиславович – доктор медицины, профессор Университета г. Тромсё, Норвегия; старший советник Норвежского института общественного здравоохранения, г. Осло, Норвегия; Директор Архангельской международной школы общественного здоровья ГБУО ВПО «Северный государственный медицинский университет» Министерства здравоохранения Российской Федерации, г. Архангельск

Адрес: Nasjonalt folkehelseinstitutt, Pb 4404 Nydalen, 0403 Oslo, Norway

Тел.: +47 22048319, +47 45268913; e-mail: angr@fhi.no