Тема 4. ВВЕДЕНИЕ В БИОСТАТИСТИКУ. ОПИСАНИЕ ДАННЫХ.

В медицине проведение клинических исследований напрямую связано с всесторонним анализом полученных данных. Поэтому изучение прикладной статистики является неотъемлемой частью обучения персонала, принимающего участие не только в статистическом анализе результатов, но и в процессе сбора клинических данных. Этические и экономические соображения диктуют необходимость внимательного отношения к планированию клинических исследований. Кроме того, владение методиками обработки информации позволяет персоналу более эффективно организовать процедуру сбора исходных данных.

Биометрия, или **Биологическая статистика (Биостатистика)** — научная отрасль на стыке биологии и вариационной статистики, связанная с разработкой и использованием статистических методов в научных исследованиях (как при планировании количественных экспериментов, так и при обработке экспериментальных данных и наблюдений) в биологии, медицине, здравоохранении и эпидемиологии [1].

В здравоохранении и клинической медицине часто используются, сознательно или неосознанно, различные статистические концепции при принятии решений по таким вопросам, как клинический диагноз, прогнозирование возможных результатов осуществления тех или иных программ в данной группе населения, прогнозирование течения заболевания у отдельного больного, выбор лечения для конкретного больного, и т.п. Статистика находит повседневное применение в лабораторной практике. Знание статистики стало важным для понимания и критической оценки сообщений в медицинских журналах. Таким образом, знание принципов статистики абсолютно необходимо для планирования, проведения и анализа исследований, посвященных оценке различных ситуаций и тенденций в здравоохранении, а также для выполнения научных исследований в области медицинской биологии, клиники и здравоохранения.

Применение статистики в здравоохранении необходимо как на уровне сообщества, так и на уровне отдельных пациентов. Медицина имеет дело с индивидуумами, которые отличаются друг от друга по множеству характеристик, таких, как масса тела, возраст, рост, артериальное давление, уровень холестерина, иммуноглобулинов и т.д. Значения показателей, на основании которых человека можно считать здоровым, варьируются от одного индивидуума к другому. Нет двух совершенно одинаковых пациентов или двух групп индивидуумов, однако решения, касающиеся отдельных больных или групп населения, приходится принимать, исходя из опыта, накопленного в отношении других больных или популяционных групп со сходными биологическими и социальными характеристиками. Ввиду существующих различий эти решения не могут быть абсолютно точными – они всегда сопряжены с некоторой неопределенностью. В этом и заключается вероятностная природа медицины.

Вариация признака (или фактора, или результатов измерения) возникает, если их значения меняются от индивидуума к индивидууму или для одного индивидуума во времени. Едва ли не всем характеристикам организма человека свойственна вариабельность. Кроме того, возникает проблема обработки результатов очень большого числа измерений. Например, если бы можно было изучить всех больных туберкулезом в мире, то такая группа больных составила бы генеральную совокупность.

Статистическая совокупность – группа, состоящая из большого числа относительно однородных элементов (объектов), взятых вместе в известных границах времени или пространства

Генеральная совокупность состоит из всех единиц наблюдения, которые могут быть к ней отнесены в соответствии с целью исследования. Естественно, практически это невозможно, поэтому при изучении здоровья населения генеральная совокупность рассматривается в пределах конкретных границ, очерченных территориальным или

производственным признаком, и поэтому включает в себя определенное число наблюдений.

Выборочная совокупность (выборка) – часть генеральной совокупности, по свойствам которой судят о генеральной совокупности. На основе анализа выборочной совокупности можно получить достаточно полное представление о закономерностях, присущих всей генеральной совокупности. Выборочная совокупность должна быть репрезентативной, т.е. в отобранной части должны быть представлены все элементы в том соотношении, как и в генеральной совокупности. Выборочная совокупность должна отражать свойства генеральной совокупности, т.е. правильно ее представлять.

Поскольку обычно имеется совокупность наблюдений (десятки, сотни, а иногда – тысячи результатов измерений индивидуальных характеристик), то возникает задача компактного описания имеющихся данных. Для этого используют **методы описательной статистики.**

Все изучаемые показатели варьируются, но не все они поддаются непосредственному измерению. Так возникает деление на *количественные* и *качественные* показатели. Классификация типов данных (рис. 4.1) приведена согласно [7].

Для дальнейшего корректного применения статистических методов необходимо понимать, в какой шкале представлены данные.

Различают следующие типы шкал:

· ***Номинальная (номинативная) шкала.***

· ***Порядковая (ординальная) или ранговая шкала.***

· ***Интервальная шкала.***

· ***Шкала отношений (шкала отношений или равных отношений).***

Первые два вида шкал называются неметрическими (используются для качественных данных), а последние два – метрическими шкалами (для количественных данных).

Шкалы измерений определяют допустимые для данной шкалы математические преобразования, а также типы отношений, отображаемых соответствующей шкалой.

Рис. 4.1. Типы медико-биологических данных, используемых в статистическом анализе.

Описательная (дескриптивная) статистика используется для первичной обработки данных, т.е. их обобщения, систематизации, наглядного представления в форме графиков и таблиц, а также их количественного описания посредством основных статистических показателей.

В зависимости от типа данных используются те или иные описательные статистики.

Описание количественных данных

Описание начинают с графического представления данных выборочной совокупности.

Статистическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот [2]

Пусть известно статистическое распределение частот количественного признака *X*.

Эмпирическим распределением (эмпирической функцией распределения) называют распределение (функцию распределения) выборочной совокупности.

Для наглядности эмпирическое распределение представляют в виде *гистограммы*.

Построение гистограмм используется для получения эмпирической оценки плотности распределения случайной величины. Для построения гистограммы наблюдаемый диапазон изменения случайной величины разбивается на несколько интервалов и подсчитывается доля от всех измерений, попавшая в каждый из интервалов. Величина каждой доли, отнесенная к величине интервала, принимается в качестве оценки значения плотности распределения на соответствующем интервале.

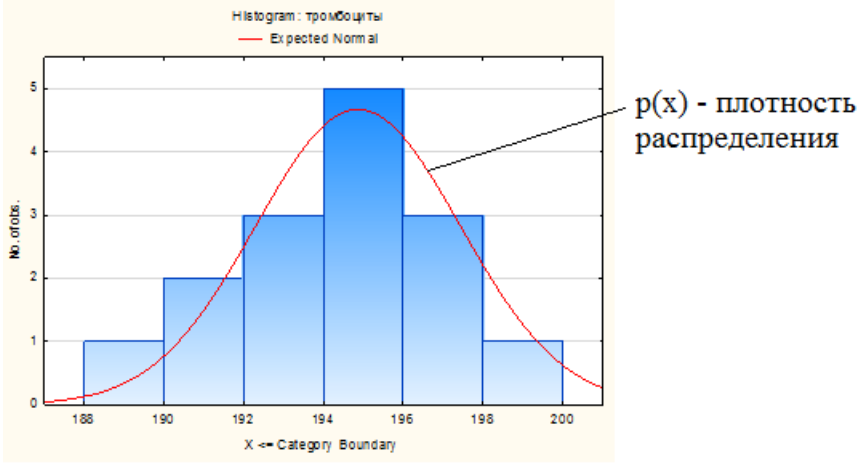

Рис. 4.2. Пример гистограммы

Основные статистические показатели можно разделить на две группы: ***меры положения*** и ***меры рассеяния***.

Меры положения – это общее понятие для числового выражения локализации данных (на числовой оси) как типичного результата измерения. Самыми распространенными из них являются среднее и медиана.

Среднее арифметическое, которое очень часто называют просто «среднее», получают путем сложения всех значений и деления этой суммы на число значений в наборе. Это можно показать с помощью алгебраической формулы. Набор *n* наблюдений переменной *x* можно изобразить как $\{x_1; x_2; \ldots; x_n\}$. В таком случае формула для определения среднего арифметического наблюдений μ имеет вид (4.1):

$$\mu = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (4.1).$$

Например, серия наблюдений (статистическая совокупность) артериального систолического давления в 11-ти наблюдениях имеет следующий вид:

Таблица 4.1.

1	2	3	4	5	6	7	8	9	10	11
120	115	120	125	120	115	120	115	120	120	125

Среднее арифметическое значение в данном ряду будет равно $\frac{120+115+120+125+120+115+120+115+120+120+125}{11} = 119{,}5$.

Медиана. Медиана делит ряд упорядоченных значений (*вариационный ряд*) пополам с равным числом этих значений как выше, так и ниже ее (левее и правее медианы на числовой оси).

Если число наблюдений *n* нечетное, это будет наблюдение номер $\frac{n+1}{2}$. Если *n* четное, то, строго говоря, медианы нет. Однако обычно можно вычислять ее как среднее арифметическое двух соседних средних наблюдений в упорядоченном наборе данных (т. е. наблюдений номер $\frac{n}{2}$ и $\frac{n}{2}+1$).

В рассмотренном выше примере медиана равна 120.

Мода – это значение, которое встречается наиболее часто в наборе данных; если данные непрерывные, то их обычно группируют и вычисляют модальную группу. Некоторые наборы данных не имеют моды, потому что каждое значение встречается только 1 раз. Иногда бывает более одной моды; это происходит тогда, когда 2 значения или больше встречаются одинаковое число раз и встречаемость каждого из этих значений больше, чем любого другого значения. В этом случае мода совпадает с минимальным модальным значением.

Для данных из таблицы 1 мода, очевидно, равна 120.

<u>Меры рассеяния</u> **–** это статистические показатели, характеризующие степень вариации, разброса значений признака относительно среднего значения.

Размах (интервал изменения) – это разность между максимальным и минимальным значениями переменной в наборе данных.

Расположим данные, полученные в таблице 4.1, упорядоченно:

Таблица 4 . 2

1	2	3	4	5	6	7	8	9	10	11
115	115	115	120	120	120	120	120	120	125	125

$размах = 115 - 125 = 10$.

Размах, полученный из процентилей. Предположим, что данные расположены упорядоченно от самой маленькой величины и до самой большой величины. Величина *X*, до которой расположен 1% наблюдений (и выше которой расположены 99% наблюдений), называется первым *процентилем*. Величина *X*, до которой находится 2% наблюдений, называется 2-м процентилем, и т.д. Величины *X*, которые делят упорядоченный набор значений на 10 равных групп, т. е. 10-й, 20-й, 30-й,..., 90 и процентили, называются *децилями*. Величины *X*, которые делят упорядоченный набор значений на 4 равные группы, т.е. 25-й, 50-й и 75-й процентили, называются *квартилями*. 50-й процентиль – это медиана.

Ряд из таблицы 5.2 можно охарактеризовать так: I квартиль (25 процентиль)=115, II квартиль (50 процентиль, медиана) = 120, III квартиль (75 процентиль)=120 (Рис. 4.3).

Рис. 4.3. Квартили и медиана в ряду измерений.

Дисперсия – мера разброса (рассеяния) данной случайной величины.

Если имеется n наблюдений $\{x_1; x_2; ...; x_n\}$, μ среднее арифметическое, то дисперсия рассчитывается по формуле (4.2):

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n - 1} \qquad (4.2)$$

Стандартное отклонение. Стандартное (среднеквадратичное) отклонение (σ) – это положительный квадратный корень из дисперсии. Оно вычисляется в тех же единицах (размерностях), что и исходные данные и характеризует степень рассеивания вариационного ряда вокруг средней.

На практике часто приходится сравнивать изменчивость признаков, выраженных разными единицами, например, рост в см и масса в кг. Если разделить стандартное отклонение на среднее арифметическое и выразить результат в процентах, получится ***коэффициент вариации***. Он является мерой рассеяния, не зависящей от единиц измерения (безразмерной) (4.3).

$$C_v = \frac{\sigma}{\mu} \times 100\% \qquad (4.3).$$

При $C_V < 10\%$ наблюдается слабое разнообразие признака, при $10\% < C_v < 20\%$ – среднее разнообразие признака, при $C_V > 20\%$ – сильное разнообразие признака.

<u>Понятие вероятности.</u> Вероятность того или иного события при числе наблюдений N оценивается по простой формуле. **Если число наблюдаемых конкретных событий при числе наблюдений N равно n, то вероятность равна отношению числа наблюдений, в которых было обнаружено событие к общему числу наблюдений** (5.5)**:**

$$P(A) = \frac{n}{N} \qquad (4.4)$$

Вероятность можно оценить в непрерывной шкале от 0 до 1 включительно. Событие, которое невозможно, имеет вероятность 0, а событие, которое произойдет обязательно, имеет вероятность 1.

<u>Математическое ожидание.</u> Пусть определена совокупность измерений систолического давления у некоторой группы обследуемых (табл. 4.2).

Что можно сказать о величине АД в следующем, двенадцатом наблюдении, которое мы не проводили? В полной мере оценить эту величину мы не можем, а лишь дать вероятностную оценку, т.е. предсказать значение с той или иной долей вероятности.

Любое измеренное нами значение АД является случайной величиной. Если имеется какая-либо зависимость, описывающая эту случайную величину, то принято говорить, что случайная величина характеризуется функцией вероятности. В этом случае, основываясь на полученных результатах, можно прогнозировать ту величину, которая будет получена в следующих измерениях. Такая прогнозируемая величина называется *математическим ожиданием***.** Попытаемся определить величину математического ожидания для нашего случая.

Для этого вначале сгруппируем одинаковые результаты и оценим вероятность (в долях единицы) их появления в нашем наблюдении (табл. 4.3):

Таблица 4.3

Систолическое АД (X)	число пациентов	вероятность (P)
115	3	3/11
120	6	6/11
125	2	2/11

Так как общее число наблюдений составило 11, то каждое появление того или иного результата представляет собой вероятность, равную 1/11.

Математическое ожидание ($M_{f(x)}$) вычисляется по следующей формуле (4.5):

$$M_{f(x)} = X_1 \cdot p_1 + X_2 \cdot p_2 + \cdots + X_n \cdot p_n \qquad (4.5)\ .$$

Математическое ожидание - это сумма попарных произведений наблюдаемой величины X_i на вероятность ее появления p_i в данном наблюдении.

В рассмотренном нами случае вариационного ряда систолического давления математическое ожидание исследуемой величины составляет:

$M_{f(x)} = 115 \cdot 3/11 + 120 \cdot 6/11 + 125 \cdot 2/11 \approx 119{,}55$.

Таким образом, наиболее вероятной будет величина, составляющая 119,55 мм рт. ст.

Распределение вероятности. *Случайная переменная* – это величина, которая может принимать любое из набора взаимоисключающих значений с определенной вероятностью. Распределение вероятности показывает вероятности всех возможных значений случайной переменной. Это ***теоретическое распределение***, которое выражено математически и имеет среднее и дисперсию – аналоги среднего и дисперсии в эмпирическом распределении. Каждое распределение вероятности определяется некоторыми параметрами. Параметры служат обобщающими величинами (например: среднее, дисперсия), характеризующими данное распределение (т.е. их знание позволит подробно описать распределение) С помощью соответствующей статистики можно произвести оценку этих параметров в выборке. В зависимости от того, является ли случайная переменная дискретной или непрерывной, распределение вероятности может быть либо дискретным, либо непрерывным.

Функция *F(x)*, связывающая значения x_i переменной случайной величины X с их вероятностями p_i называется ***законом распределения*** (или ***функцией распределения***) этой случайной величины. Закон распределения описывает распределение вероятностей случайной переменной *X*.

С понятием закона распределения случайной величины неразрывно связано понятие ***плотности распределения***, которую можно представить себе как предельную кривую *p(x)*, аппроксимирующую выборочную гистограмму распределения данной случайной величины (рис. 4,1, 4.4).

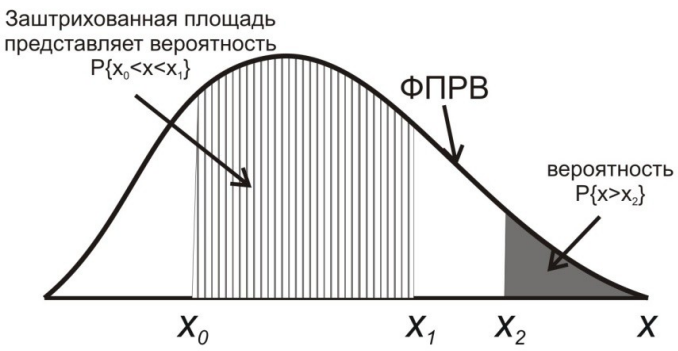

Рис. 4.4. Функция плотности распределения вероятности.

Нормальное (гауссово) распределение. Одним из самых важных распределений в статистике является нормальное распределение.

Непрерывная случайная величина X называется распределенной по нормальному закону, если ее плотность распределения равна

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad (4.6)$$

где m - математическое ожидание случайной величины;

Его функция плотности распределения вероятности представлена на рис. 4.5.

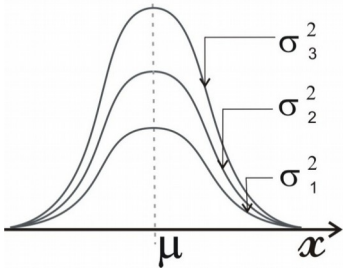

Рис. 4.5. Функция плотности нормального распределения вероятности.

Функция плотности нормального распределения вероятности симметрична относительно среднего μ. Результат изменения σ^2 ($\sigma_1^2 > \sigma_2^2 > \sigma_3^2$).

Свойства функции плотности нормального распределения вероятности:

• полностью определяется двумя параметрами – средним (μ) и дисперсией (σ^2);

• колоколообразная (унимодальная) форма;

• симметрична относительно среднего;

• сдвигается вправо, если среднее увеличивается, и влево, если среднее уменьшается (при постоянной дисперсии);

• сплющивается, если дисперсия увеличивается, но становится более остроконечной, если дисперсия уменьшается (для постоянного среднего).

• среднее и медиана нормального распределения равны

• вероятность того, что нормально распределенная случайная величина X со средним арифметическим μ и стандартным отклонением σ попадает в интервалы $\pm 1\sigma$, 2σ и 3σ равны соответственно

(μ - σ) and (μ + σ) 0,68

(μ -1,96σ) and (μ + 1,96σ) 0,95

(μ -2,58σ) and (μ + 2,58σ) 0,99

Нормальное распределение не является единственным известным распределением.

Примеры распределений непрерывных и дискретных случайных величин: t-распределение (распределение Стьюдента), Хи-квадрат распределение Пирсона, F-распределение (распределение Фишера), логнормальное распределение, биномиальное распределение, распределение Пуассона.

Стандартная ошибка среднего. Случайные ошибки выборок возникают за счет того, что для анализа всей совокупности используется только ее часть. Хотя выборочный метод и позволяет обоснованно судить о средней арифметической некоторого количественного признака генеральной совокупности по средней арифметической, исчисленной по выборке, это, однако, не означает, что выборочная средняя совпадает с генеральной средней. Она, как правило, в той или иной степени от нее отличается. Величина ошибки выборки представляет собой разность между генеральной и выборочной средними. Ошибки выборки различны для каждой конкретной выборки и в принципе могут быть обобщенно охарактеризованы с помощью средней из всех таких отдельных ошибок. В математической статистике получены формулы, которые позволяют приближенно вычислить среднюю ошибку выборки, основываясь на данных только той выборки, которая имеется в распоряжении исследователя.

Стандартная ошибка среднего отражает точность оценки среднего значения признака в популяции по его выборке. Небольшая стандартная ошибка (существенно меньше соответствующего среднего значения) означает достаточно точную оценку. Стандартная ошибка уменьшится, т. е. оценка станет более точной, если объем выборки увеличится или данные имеют небольшое рассеяние (дисперсию). При неограниченном увеличении объема выборки стандартная ошибка среднего обращается в 0.

Стандартная ошибка среднего арифметического может быть найдена по формуле (4.7):

$$\sigma(x) = \frac{\sigma}{\sqrt{n}}, \quad (4.7)$$

где σ – среднее квадратическое отклонение,

n – количество параметров в выборочной совокупности.

Доверительный интервал. Выборка из популяции позволяет получить точечную оценку интересующего нас параметра и вычислить стандартную ошибку для того, чтобы указать точность оценки. Следует отметить, что для большинства исследований стандартная ошибка как таковая неприемлема, поскольку она, в отличие от стандартного отклонения, не отражает вариабельности в значениях данных. Гораздо полезнее объединить эту меру точности с **интервальной оценкой** для параметра популяции. Для этого нужно вычислить **доверительный интервал** (ДИ), который дает вероятное значение верхней и нижней границ оцениваемой неизвестной величины, что позволяет заявить: «Я утверждаю, что точное значение неизвестной величины с определённой вероятностью (чаще всего эта вероятность составляет 0,95) находится между этими двумя числами».

В случае нормального распределения доверительный интервал среднего значения - $(\mu - tm; \mu + tm)$, где t – коэффициент Стьюдента – величина, зависящая от объема выборки (или соответствующего числа степеней свободы) и выбранного уровня доверительной вероятности, определяется по таблицам распределения Стьюдента, а m – стандартная ошибка среднего.

В случае распределения, отличного от нормального, вычисляют медиану x_{50}, квартили (x_{25}, x_{75}) и статистически значимый диапазон — например:

$$x_1 = \max(x_{min}, x_{25} - 1{,}5 \cdot (x_{75} - x_{25}))$$
$$x_2 = \min(x_{max}, x_{75} + 1{,}5 \cdot (x_{75} - x_{25})).$$

При интерпретации ДИ исследователь формулирует следующие вопросы:

1. Насколько широк ДИ? Широкий ДИ указывает на менее точную оценку, узкий - на более точную оценку.
2. Какой клинический (биологический) смысл можно извлечь из рассмотрения ДИ? Верхние и нижние пределы показывают, будут ли результаты клинически (биологически) значимы.
3. Включает ли ДИ какие-либо значения, представляющие особый интерес?

Описание качественных данных

Единственный способ описания качественных признаков состоит в том, чтобы подсчитать число объектов, обладающих одним и тем же качественным признаком, или долю объектов с одним и тем же признаком от общего числа объектов.

В отношении оценки долей возникают те же статистические задачи, что и для параметров, представленных в количественной форме:

- оценка доли p в генеральной совокупности по выборочным данным, нахождение доверительного интервала для p;
- выявления различия между генеральными долями p1 и p2 двух совокупностей по выборочным данным, т.е. сравнение двух выборочных долей вариант.

Ошибка доли рассчитывается с помощью формулы (4.8)

$$s_p = \sqrt{\frac{p(1-p)}{N}}, \quad (4.8),$$

где s_p – ошибка доли,

p – доля признака в выборке.

95% доверительный интервал для доли признака в генеральной совокупности будет представлен в виде $p \pm 1{,}96 s_p$

Большинство критериев и статистических тестов относятся к так называемым *параметрическим критериям.* Это значит, что они могут применяться **только** к нормально распределенным рядам данных. Во всех остальных случаях используются так называемые *непараметрические критерии.* В случае, когда распределение ряда параметров является отличным от нормального или о природе распределения ничего не известно, необходимо обращаться именно к таким методам. Говоря более специальным языком, непараметрические методы не основываются на оценке параметров (таких как среднее или стандартное отклонение) при описании выборочного распределения интересующей величины. Поэтому эти методы иногда также называются свободными от параметров или свободно распределенными.

Если данные не являются нормально распределенными, а измерения, в лучшем случае, содержат ранжированную информацию, то вычисление обычных описательных статистик, таких например, как среднее и стандартное отклонение, не слишком информативно.

Непараметрическая статистика вычисляет такие параметры, как медиана, мода, интерквартильный размах и др., позволяющие получить более "полную картину" данных.

Таким образом, статистический анализ медико-биологических данных должен начинаться с их первичной обработки, т.е. представления исходных данных в подходящей для анализа форме, и проведения проверки качества данных.

Порядок первичной обработки данных (предварительный анализ данных) представлен на рис 4.6.

Рис 4.6. Первичная обработка (предварительный анализ)данных.

Контрольные вопросы

1. Для чего может применяться математическая статистика в медицине?
2. Перечислите основные описательные статистики, используемые в предварительном анализе данных.
3. Какие меры положения и меры рассеяния признака вам известны?
4. Что отражает стандартная ошибка среднего?
5. Что такое закон распределения случайной величины?

6. В чем разница между параметрическими и непараметрическими критериями?

Список литературы

1. Лакин Г. Ф. Биометрия: Учебное пособие для вузов – 4-е изд., перераб. и доп. – М.: Высш. шк., 1990. – 352 с.
2. Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. пособие для вузов – 9-е изд., стер. – М.: Высшая школа, 2003. – 479 с.
3. Лях Ю.Е., Гурьянов В.Г., Хоменко В.Н., Панченко О.А. Основы компьютерной биостатистики: анализ информации в биологии, медицине и фармации статистическим пакетом Medstat. – Донецк:, 2006. – 214 с.
4. Островок здоровья. – Режим доступа: www.bono-esse.ru
5. Петри А., Сэбин К. Наглядная статистика в медицине. – М.: ГЭОТАР-МЕД, 2003. – 139 с.
6. Платонов А.Е. Статистический анализ в медицине и биологии: задача, терминология, логика, компьютерные методы. – М.: Издательство РАМН, 2000. – 52 с.