**КОМПЬЮТЕРНЫЙ АНАЛИЗ МЕДИЦИНСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ**

**Программные средства математической статистики**

Математическая статистика — универсальный инструмент для анализа любых данных, в том числе экспериментальных клини­ческих и биомедицинских. Но выбираемый метод должен отве­чать поставленной цели и быть адекватным по отношению к ха­рактеру анализируемых данных. Современный врач-исследователь должен осмысленно выбирать методы, применяемые к конкрет­ной клинической (экспериментальной) задаче, и критически оценивать, а также содержательно интерпретировать получен­ные результаты.

Статистический анализ можно проводить вручную, если дан­ных немного, используемые методы просты, а расчеты вследствие этого не трудоемки. Но в подавляющем большинстве случаев не­обходимо пользоваться специальными программными пакетами для ПК, которые так и называются — статистическими.

С конца 1970-х гг. в России самым популярным для использова­ния в медицине и биологии был пакет *BMDP.* Обработка данных проводилась на так называемых «больших» машинах серии ЕС с предварительной набивкой данных на перфокарты.

В первой половине 1990-х гг. лидерство захватили статистиче­ские пакеты для персональных ЭВМ, работающие под управле­нием ОС *MS DOS.* Одним из математически мощных, не наклады­вающих практически никаких ограничений на объем обрабатыва­емой информации, был пакет *SAS.* Часть исследователей работала с пакетом *BMDP* для ПК, но безусловным лидером по количеству пользователей был пакет *Statgraphics*, обладающий широкими воз­можностями, достаточно простой в эксплуатации, но имеющий ограничения по числу анализируемых переменных.

В настоящее время наибольшее распространение в России по­лучили статистические пакеты, работающие под *Windows: Statistica, SPSS, SAS*

Для того чтобы успешно применять математическую статисти­ку, ее нужно знать хотя бы в том объеме, который реально необходим для осознанных действий. Кроме того, нужно уметь исполь­зовать статистический пакет. С практической точки зрения лучше ориентироваться на один и тот же пакет в течение ряда лет. Пере­ход от пакета к пакету ведет к необходимости переучивания, пусть и в относительно небольшом объеме.

**3.2. Особенности медицинских данных**

Первым шагом, предваряющим собственно статистический анализ, является исследование типа данных, основными из кото­рых являются количественные и качественные.

Качественные данные подразделяются на порядковые, или ран­говые (например, тяжесть проявлений заболевания), и классифи­кационные, или номинальные (например, перенесенные заболе­вания, классы ксенобиотиков). Процедура ранжирования данных, т.е. упорядочивания их в соответствии с числовыми градациями, проводится в возрастающем, либо в нисходящем ряду значений. Число градаций, характеризующих данные, не должно быть из­лишне большим, так как в этом случае увеличивается элемент субъективности.

При обработке данных часто приходится переводить качествен­ные данные в количественные. В свою очередь количественные данные могут подвергаться квантованию в зависимости от постав­ленной задачи (для выделения интервалов, соответствующих раз­личным состояниям, например температура нормальная, субфебрильная, высокая и др.), и тогда они становятся аналогичны ка­чественным шкалированным.

В случае пропусков информации (отсутствующие данные) нельзя использовать так называемое «обнуление», т.е. приписы­вать кодовое число нуль, так как это в большинстве случаев со­впадает с кодированием нормы по данному признаку. Также ме­тодически неверно использовать среднее по классу, особенно при малых выборках, так как классы далеко не всегда являются одно­родными (гомогенными). Предпочтительно исключать такие на­блюдения или кодировать пропущенные данные специальным знаком (числом) при условии, что «обход» таких «значений» пред­усмотрен в программе, т.е. обработка проводится только по изве­стным значениям данных.

В клинической и экспериментальной медицинской практике исследователь реже употребляет слово «данные», но чаще — *«па­раметры»* или «*переменные*», ставя между этими понятиями знак равенства.

Нужно отличать понятие *«переменная»* от понятия *«признак»:* температура тела — параметр (переменная), температура тела бо­лее 37° С — признак (человек нездоров). Переменные бывают непрерывными и дискретными, в том числе, дихотомическими (принимающими одно из двух значений, например «здоров — болен»).

В клинической практике переменные часто описываются шка­лами. Как было отмечено ранее, шкалы бывают качественными (сознание ясное, спутанное и т.д.), классификационными (цвет кожных покровов — розовые, синюшные, желтушные, характер хрипов в легких — сухие, влажные мелкопузырчатые, крупнопу­зырчатые и т.д.) и количественные, в том числе интервальные, порядковые, балльные.

Интервальные переменные (например, ударный объем, мл: 50 — 80, менее 50, более 80) полезно использовать для решения конк­ретной клинической задачи. Их также можно переводить в поряд­ковые (на основе построения шкал), но лучше (при наличии воз­можности и если это не противоречит смыслу решаемой задачи) использовать собственно количественные значения переменной.

Балльные шкалы получают по-разному: после предваритель­ной математической обработки, на основе чисто клинической оценки параметра, комбинированным способом.

Как правило, затруднения в отнесении параметров к количе­ственным или качественным не возникают. Гораздо сложнее при переводе качественных данных в количественные, когда реально нужно определить «вес» градаций качественного параметра. Не всегда правомерен вариант «в лоб»: 1, 2, 3, 4 и т.д. (но не более девяти значений, исходя из известного закона в области психоло­гии, определяющего пределы способности человека к переработ­ке информации). Часто для получения реальных «весов» парамет­ров необходимо проводить дополнительное исследование.

Особым типом данных являются даты. Бывает, что по смыслу работы с ними приходится производить действия (например, оп­ределять, сколько времени прошло между соседними исследова­ниями), поэтому нужно следить, чтобы они записывались в опре­деленном формате.

Важен вопрос о точности измерения и представления меди­цинских параметров. Понятно, что точность исходных данных оп­ределяется точностью метода и(или) прибора, с помощью кото­рых осуществляется измерение.

В *описательной статистике* при работе с медицинскими дан­ными необходимо поступать следующим образом: с одной сторо­ны, не допускать потерь информации исходно — использовать данные с той точностью, которая имеет место при измерении; с другой — при представлении результатов статистической обра­ботки данных не приводить избыточной информации — в боль­шинстве случаев достаточно той точности представления инфор­мации, что и в исходных данных, либо использования одного Дополнительного разряда. Обычно при предъявлении числовых данных указываются два знака после запятой. Исключением являются случаи представле­ния констант и весовых коэффициентов функций, полученных в результате многомерного анализа (например, дискриминантного); тогда в результирующих таблицах обязательно должны приводиться все цифры после запятой.

Следует остановиться на еще одной особенности медицинских данных.

В *математической статистике* выводы основаны на допуще­нии: то, что верно на случайной выборке, верно и для генераль­ной совокупности, из которой она получена.

*Генеральная совокупность —* это набор данных, описывающих нечто всеобъемлющее, например все дети, живущие на планете Земля, вся совокупность пациентов, которые могли бы получать определенный пре­парат и т. п.

*Выборка* — часть генеральной совокупности, описывающая ее с той или иной долей погрешности.

Часто сформировать истинно случайную выборку из генераль­ной совокупности не представляется возможным в силу того, что для выполнения требований репрезентативности объекты иссле­дования (пациенты) должны проживать на разных континентах земного шара. Проведение таких исследований в настоящее время возможно, однако в силу существенных физиологических разли­чий между людьми, живущими в разных уголках планеты, может быть потерян клинический смысл исследования. Это утверждение справедливо для достаточно широко распространенных в насто­ящее время международных мультицентровых исследований, по­священных, например, метаболическому синдрому, в которых принимают участие крупные отечественные центры и институты. В таких случаях выборка должна быть репрезентативна к исследу­емой популяции (населению РФ или определенных, этнически сходных, регионов России).

Современная технология статистического анализа данных вклю­чает:

1) постановку задачи и планирование исследования: составле­ние детального плана сбора исходных данных, определение ха­рактера выборки;

2) подготовку данных;

3) выбор методов обработки данных;

4) проведение анализа данных;

5) интерпретацию и представление результатов анализа.

Суть современной технологии обработки медицинских данныхс помощью методов математической статистики и их последу­ющей интерпретации кратко изложена в подразд. 3.3 — 3.5.

**3.3. Подготовка, предварительный анализ информации и выбор методов обработки данных**

Рассмотрим пример из клинической дисциплины.

**Постановка задач и планирование исследования.** Предпочтитель­ным вариантом является строгий подход, когда до проведения исследования есть полная ясность, для чего предпринимается ра­бота, сколько и каких исследований будет осуществлено, какие методы и почему будут применяться для обработки полученного материала. Это позволяет оптимизировать усилия исследователя и затраты ресурсов.

Данный вариант исследования не исключает использования (целиком или частично) ретроспективных данных (собранных ранее и имеющихся в медицинской документации). В принципе нужно помнить, что в этом случае исследователя подстерегает еще одна «ловушка», связанная с различным толкованием симптомов. Это же может иметь место и при проспективном исследовании с участием группы врачей, особенно придерживающихся различ­ных научных школ. Поэтому помимо списка признаков, характе­ризующих заболевание, необходимо иметь их стандартные описа­ния. Это в значительной степени позволит избежать различий в толковании одних и тех же симптомов. Однако в клинической прак­тике до сих пор достаточно часто сначала получают данные, а уже затем клиницист решает, каким образом на имеющемся материа­ле получить некий обобщающий результат.

**Подготовка данных.** Данные для статистического анализа при­нято готовить в виде таблицы (таблиц). Современные статистиче­ские пакеты работают с данными наиболее распространенных в настоящее время форматов, в том числе *.dbf* и *.xls.*

В строки таблицы заносятся объекты исследования (например, паци­енты), а в столбцы — параметры. Если конкретное значение параметра отсутствует, клетку таблицы оставляют пустой. Если значение параметра равно нулю, оно все равно обязательно вносится. Если пациенты обсле­довались в динамике, т.е. по каждому больному есть несколько «срезов» параметров, обычно вводится дополнительный столбец, однозначно определяющий для конкретного больного (и соответственно — значений параметров) время исследования (например, номер хирургических су­ток). Таблицы данных включают в себя столбец (столбцы) группирующих параметров (например, номер группы, исход заболевания, если именно по нему будут исследоваться пациенты, и т.д.).

После занесения данных в таблицу необходимо их проверить: просматривают значения параметров, выявляют те из них, кото­рые сильно отличаются от остальных. Это могут быть как реаль­ные «выпадающие» значения, так и ошибки ввода, которые не­обходимо устранить.

Перенос (импорт) данных в статистический пакет затрудне­ний не вызывает. Для этого пользуются стандартными возможно­стями буфера обмена ОС *Windows.* Можно также применять специ­альные модули статистических пакетов, например модуль «Уп­равление данными» (*Data management)* пакета *Statistica.*

Современные статистические пакеты дают возможность управ­лять данными: часто при решении задач возникает необходимость объединения или разделения файлов (содержащих таблицы) по условию.

**Выбор методов анализа и их реализация.** Для грамотного выбо­ра метода обработки данных необходимо знать характер распреде­ления используемых переменных, поэтому предварительный ана­лиз данных начинают с определения характера их распределения.

*Распределение элементов выборки по значениям параметра* — это сово­купность частот встречаемости интервалов его значений в данной вы­борке. К наиболее часто встречающимся видам распределений относят­ся: колоколообразное (нормальное, гауссово), полимодальное (чаще — бимодальное), равномерное и др.

К основным характеристикам распределения относятся:

* среднее арифметическое (М) — при непрерывных числовых типах параметров; все значения по выборке сложить и поделить на их количе­ство;
* медиана — значение параметра, делящее распределение параметра пополам; выборка значений параметра ранжируется (по возрастанию или убыванию); если число значений нечетно, то медиана — это централь­ное значение, если число значений четно, то медиана — это среднее арифметическое двух центральных значений;
* квантили (центили) — весь диапазон значений разбивается на 10 интервалов. Границы между интервалами — квантили, от 10%-го до 90%-го квантиля укладываются 80 % значений;
* квартили — весь диапазон разбивается на четыре интервала: 50%-й квартиль — медиана, кроме нее часто указываются 25%-й и 75%-й квар­тили, т.е. описывается 50 % наиболее «близких к центру» значений;
* мода — значение параметра с наибольшей частотой встречаемости на выборке;
* асимметрия — характеристика несимметричности распределения элементов выборки относительно среднего арифметического. В случае симметричного распределения значение асимметрии равно нулю.

В медицинских публикациях часто встречается запись значений в виде *М± т,* где *т* — стандартная ошибка среднего (*standard error of mean).* Это допустимо делать в случае нормально распределен­ного параметра, а к величине т нужно относиться с определен­ной долей скептицизма. Правда, при увеличении выборки рас­пределение параметра достаточно часто стремится к квазинормаль­ному, и тогда использование т в какой-то мере оправдано. Лучше указывать само выборочное стандартное отклонение (среднее квад­ратичное отклонение — *standard deviation — s*), которое характе­ризует ширину нормального распределения. Основанием для та­кого подхода является то, что s не уменьшается при увеличении числа наблюдений *п;* в диапазон *М ± s* укладывается около 70% значений нормального распределения параметра.

*Параметрические методы.* Для решения многих клинико-научных задач необходимо формулировать статистические гипотезы. Среди них можно назвать анализ соответствия распределения зна­чений параметра определенному закону, сравнение групп по ха­рактеристикам распределения параметров и др.

Статистическая гипотеза — это формально строго сформули­рованное предположение.

Нулевой (Н0) называют гипотезу, которую исследователь предпола­гает отклонить (например, об отсутствии различий между группами).

Альтернативная гипотеза(Н1) противоположна нулевой (например, о наличии различий между группами).

Уровень статистической значимости (а) — это пороговое значение для ошибочного отклонения верной нулевой гипотезы (ошибки первого рода). В медицине принято выбирать а = 0,05 или а = 0,01.

Ошибка второго рода — это ошибочное принятие ложной нулевой гипотезы.

В настоящее время в публикациях принято указывать реальное значе­ние р (вероятность ошибки первого рода). Если значение р меньше 0,05, говорят о наличии статистически значимых отличий между выборками параметра.

Статистически значимые различия следует отличать от клини­чески значимых. Встречаются результаты, значимые статистически, но не значимые с клинической точки зрения, бывает и наоборот. Клинически значимые, но статистически незначимые результаты обычно получаются на малых выборках, а при увеличении выбо­рок они, как правило, подтверждаются и статистически.

Чем распределения отличаются с практической точки зрения? Тем, что наиболее распространенные методы параметрической статистики (например, t-критерий Стьюдента) можно применять только для нормально распределенных величин (колоколообразных распределений). Неправомочное использование t-критерия Стьюдента — самая часто встречающаяся ошибка статистической обработки данных клинических исследований, приводящая к оши­бочным выводам.

*Непараметрические методы.* В клинической медицине и при об­работке данных медико-биологических экспериментов в большин­стве случаев необходимо пользоваться непараметрическими ме­тодами статистического анализа. Они являются менее мощными, чем параметрические, но применимы для любых видов распределений.

Анализ характера распределения данных (его еще называют проверкой на нормальность распределения) осуществляется по каждому параметру. Для проверки на нормальность используют как визуализирующие методы (метод построения гистограмм), так и статистические (например, тест Колмогорова—Смирнова, кри­терий Шапиро—Уилкса). Для того чтобы уверенно судить о соот­ветствии распределения параметра нормальному закону, необхо­димо, чтобы выборка была достаточно многочисленной (не менее 50 значений).

Кроме разделения по уже описанному важнейшему статисти­ческому подходу (параметрические, непараметрические) методы статистического анализа данных принято классифицировать не­сколькими способами:

1) по количеству одновременно анализируемых параметров (одномерные, двухмерные, многомерные или многофакторные);

2) имеющимся исходно предположениям о характере распре­делений выборок (односторонние тесты — при наличии предположения о смещении распределения

Таблица 1

**Методы математической статистики, используемые в клинической практике**

|  |  |
| --- | --- |
| Область применения | Метод |
| параметрический | непараметрический |
| Описательная ста­тистика | Вычисление средних значений, среднеквад­ратичных отклонений и др. | Вычисление медиан, квартилей, межквартильного размаха, квантилей и др. |
| Сравнение двух независимых групп по одному пара­метру | t-Критерий Стьюдента для независимых вы­борок | Критерий Манна-Уитни, критерий х2, точный критерий Фишера и др. |
| Сравнение двух зависимых групппо одному парамет­ру | t-Критерий Стьюдента для зависимых выбо­рок | Критерий Вилкоксона, критерий знаков и др. |
| Анализ взаимосвязи двух параметров | Корреляционный анализ по Пирсону | Корреляционный анализ по Спирмену, Кендаллу и др. |
| Одновременный анализ трех и более параметров | Регрессионный ана­лиз, дискриминантный анализ, кластер­ный анализ, диспер­сионный анализ | Логистический регрес­сионный анализ, анализ конъюнкций и др. |

параметра в одной из групп в определенную сторону относительно другой; двусторонние — при отсутствии такого предположения);

3) зависимости/независимости выборок.

Независимыми считаются, например, группы пациентов, которые были рандомизированы (случайным образом отобраны). Зависимыми являются, например, данные одной и той же группы больных до и после лечения.

Таким образом, для решения задач используют ряд параметри­ческих и непараметрических статистических методов (табл. 1).

Приведенный обзор методов, используемых при решении раз­ных задач, демонстрирует только подход к анализу данных и не претендует на полноту. Более подробно об этих методах можно узнать в соответствующей учебной литературе.

**3.4. Использование методов математической статистики для анализа данных**

В учебном издании по медицинской информатике было бы из­лишне приводить подробные описания методов математической статистики, тем более что в последние годы вышло достаточно большое количество специальной литературы, рассчитанной на практикующего врача, с описанием как наиболее часто использу­ющихся методов, так и работы со статистическими пакетами. По­этому здесь мы ограничимся лишь краткими сведениями, полез­ными для клинициста, но не достаточными для приобретения реальных знаний по математической статистике.

**Сравнение двух независимых групп по одному параметру.** t-Кри­терий Стьюдента для независимых выборок (групп) является наи­более популярным методом решения этой задачи, суть которой сводится к проверке того, различаются ли средние значения па­раметра в сравниваемых группах. Критерий корректно использо­вать только при условии нормального распределения параметров в каждой группе и равенства дисперсий распределений парамет­ров в группах.

Суть применения t-критерия Стьюдента для независимых вы­борок заключается в проверке нулевой гипотезы о том, что сред­ние значения параметра в группах не различаются. Если нулевая гипотеза по результатам анализа отклоняется *(р* < 0,05), прини­мается альтернативная гипотеза о том, что средние значения па­раметров в группах различаются.

Правомочно использовать t-критерий Стьюдента для незави­симых выборок лишь при достаточно большом объеме выборок, что в клинической медицине бывает редко.

Кроме «классического» t-критерия Стьюдента существует его модификация, не требующая равенства дисперсий распределений параметров в группах.

В настоящее время, когда врачи становятся более сведущими в математической статистике, критерий Манна-Уитни (Мапп — *Whitney U-test)* используют почти так же часто, как t-критерий. Его применяют для сравнения выборок по количественным пара­метрам в случаях, когда хотя бы одна из сопоставляемых выборок имеет распределение, отличное от нормального, или если харак­тер распределения параметра неизвестен (проверка на нормаль­ность не проводилась).

Суть метода заключается в проверке нулевой гипотезы о равен­стве средних рангов в группах, т.е. до проверки гипотезы осуще­ствляется ранжирование значений параметра в каждой группе. Если нулевая гипотеза отклоняется, принимается альтернативная ги­потеза о том, что между рангами групп есть различия.

**Сравнение двух зависимых групп по одному параметру.** t-Крите­рий Стьюдента для зависимых выборок, так же как и t-критерий Стьюдента для независимых выборок, можно применять только при условии нормального распределения параметров в каждой группе и равенства дисперсий распределений параметров в груп­пах. В большинстве случаев на реальных клинических данных эти условия не выполняются, поэтому применение метода не право­мочно.

Критерий Вилкоксона (*Wilcoxon matched pairs test) —* один из самых мощных непараметрических критериев. Его используют для парного сравнения выборок количественных (или качественных порядковых) параметров в тех случаях, когда хотя бы в одной из анализируемых выборок распределение величин параметра не яв­ляется нормальным.

При применении критерия Вилкоксона проверяется нулевая гипотеза об отсутствии различий выборок. Если она отклоняется (р < 0,05), принимается альтернативная — об их наличии.

**Анализ взаимосвязи двух параметров.** Общепринятым способом выявления взаимосвязи между переменными является расчет кор­реляции.

Следует подчеркнуть, что обнаружение корреляции между дву­мя переменными не свидетельствует о существовании причин­ной связи между ними, а лишь указывает на возможность тако­вой (или фактора, определяющего изменение обеих перемен­ных).

Обычно при использовании методов корреляции перед иссле­дователем возникает вопрос о тесноте связи (степени сопряжен­ности) переменных. Если каждому заданному значению одной переменной соответствуют близкие друг к другу, тесно располо­женные около средней величины значения другой переменной, то связь является более тесной; если эти значения сильно варьи­руют, связь менее тесная. Таким образом, мера корреляции (зна­чение коэффициента корреляции г) указывает, насколько тесно связаны между собой параметры. Чем больше коэффициент кор­реляции, тем с большей степенью уверенности можно говорить о наличии линейной зависимости между параметрами.

Условно выделяют следующие уровни корреляционной свя­зи: слабая — около 0,3; умеренная — от 0,31 до 0,5; заметная — от 0,51 до 0,7; высокая — 0,71 и более.

По форме корреляция бывает прямой (при увеличении значе­ний первой переменной значения второй также увеличиваются) и обратной (при увеличении значений первой переменной значе­ния второй убывают). Коэффициент корреляции г принимает зна­чения от -1 до +1. Обсуждать наличие корреляции имеет смысл только в тех случаях, когда она статистически значима *(р* < 0,05). Отсутствие линейной корреляции не означает, что параметры независимы: связь между ними может быть нелинейной.

Наиболее часто применяемыми в настоящее время методами исследования корреляции являются параметрический анализ по Пирсону и непараметрический анализ по Спирмену.

Корреляционный анализ по Пирсону используется при реше­нии задачи исследования линейной связи двух нормально распре­деленных параметров. Проверяется нулевая гипотеза об отсутствии связи между параметрами, т.е. что г- 0. Кроме проверки на нор­мальность распределения каждого параметра до проведения кор­реляционного анализа рекомендуется строить график в координа­тах оцениваемых параметров, чтобы визуально определить харак­тер зависимости. Если нулевая гипотеза отклоняется *(р* < 0,05), можно говорить о наличии значимой взаимосвязи между пара­метрами.

Корреляционный анализ по Спирмену применяется для ис­следования взаимосвязи двух параметров, если распределение хотя бы одного из них отлично от нормального. Проверяется нулевая гипотеза о том, что коэффициент корреляции равен нулю. Если нулевая гипотеза отклоняется *(р* < 0,05), взаимосвязь между па­раметрами есть.

**Одновременный анализ трех и более параметров.** Наряду с мето­дами одномерного и двухмерного анализа существует большое ко­личество методов многомерного (многофакторного) анализа дан­ных. Они дают возможность одновременно анализировать три и более переменные. К наиболее используемым методам многомер­ного анализа относятся: регрессионный анализ, дискриминантный анализ, кластерный анализ, дисперсионный анализ, анализ главных компонентов, факторный анализ.

В клинических работах методы многофакторного анализа ис­пользуются гораздо реже, чем описательная статистика, методы сравнения двух групп по параметру и корреляционный анализ, I хотя в последние годы наметилась тенденция к более широкому 1 применению регрессионного анализа.

*Регрессионный анализ* представляет собой метод статистического анализа, позволяющий исследовать вид зависимости одного параметра от нескольких других. Наряду с дискриминантным и кластерным он является одним из методов статистического моделирования. Моделью при этом является получаемое уравнение регрессии. С помощью рассчитываемых в ходе peгрессионного анализа константы и коэффициентов можно про­гнозировать величину исследуемого параметра в зависимости от значений других переменных. В отличие от корреляционного анализа, который лишь дает возможность установления факта взаимосвязи параметров, он описывает вид зависимости пере­менных.

Регрессионный анализ подразделяют на однофакторный (один независимый параметр) и многофакторный (два и более незави­симых параметра), а также линейный и нелинейный.

Линейный регрессионный анализ используется в тех случаях, когда все задействованные в нем параметры являются нормально распределенными, количество значений параметров намного пре­вышает количество самих параметров и т.д. Число ограничений на корректное проведение регрессионного анализа достаточно велико.

Самым употребляемым видом нелинейного регрессионного анализа в настоящее время является логистический. Главными ус­ловиями его применения является возможность принятия зависи­мым параметром только двух значений (например, есть заболева­ние — единица, нет заболевания — нуль). Все остальные парамет­ры, задействованные в анализе, должны быть независимыми, при этом они могут быть любыми по типу — как количественными, так и качественными.

*Дискриминантный анализ* — это один из методов решения зада­чи классификации — разработки правила отнесения исследуемо­го объекта к одной из нескольких групп на основании величин выделенных параметров.

*Кластерный анализ* является методом статистической группи­ровки объектов или параметров исследования в кластеры (от англ. *cluster* — гроздь, скопление) — подмножества исследуемой вы­борки.

Использование в практической деятельности врача методов многофакторного статистического анализа выходит за рамки не­обходимых знаний и навыков, которыми он должен владеть. Их применение требует глубоких знаний математической статис­тики, определенного опыта работы с медицинскими данными, а порой — даже искусства. I

**3.5. Интерпретация и представление полученных результатов**

В настоящее время написание клинического отчета, научной статьи и тем более диссертации невозможно без грамотного пред­ставления результатов, полученных с помощью методов матема­тической статистики. В работе в явном виде должна присутство­вать постановка задачи. Клинический материал представляется как фактически (количество больных, нозологические формы, возра­стной состав; данные контрольной группы), так и в отношении технологии его формирования (отбор больных и здоровых, сроки наблюдения, исследования и др.).

Если в исследовании создавались и применялись формализо­ванные карты, их вид обязательно приводят. Читателю должно быть ясно, какие параметры каждого пациента и с помощью ка­кой аппаратуры и методик получены, в каких единицах измерены и какой статистической обработке были подвергнуты.

В случае, если данные помещались в таблицы, например *MS Excel*, или создавалась БД, например *MS Access*, описывается структура таблиц. Приводится также описание всех манипуляций, которые производились с данными на предварительном этапе ис­следования.

При описании любых результатов, связанных с использовани­ем методов математической статистики, необходимо точно ука­зывать названия методов, а также название и номер версии стати­стического пакета, который применялся.

При интерпретации результатов, связанных с проверкой ста­тистических гипотез, необходимо придерживаться простого пра­вила: при р > 0,05 нулевая гипотеза не отклоняется, при р < 0,05 принимается альтернативная гипотеза о существовании различий с полученной величиной уровня статистической значимости р.

При представлении результатов применения параметрических методов статистического анализа обязательно приводятся: коли­чество объектов исследования в каждой группе, среднее и сред­неквадратичное отклонение каждого исследуемого параметра, результаты применения методов проверки на нормальность рас­пределения каждого параметра в группах, точное значение *р.*

При представлении результатов применения непараметриче­ских методов статистического анализа обязательно приводятся: количество объектов исследования в каждой группе, медианы и межквартильный размах каждого исследуемого параметра, точное значение *р.*

Для наглядности исследуемого материала при использовании как параметрических, так и непараметрических методов имеет смысл приводить данные в графической форме, демонстрирующей характер распределения величин параметров.

Представляя результаты анализа связей между параметрами любым из методов корреляционного анализа, кроме указания названия метода нужно приводить: число анализируемых пар для каждого параметра, величину коэффициента корреляции с точ­ностью до двух значащих цифр, точное значение *р.* Желательно приводить графики рассеивания объектов в координатах исследу­емых параметров.

В заключение хотелось бы еще раз подчеркнуть, что гл. 2 и 3 являются «мостиком» между основами информатики и собствен­но медицинской информатикой, в определенном смысле данью сложившейся традиции. В дальнейшем будут рассматриваться лишь особенности использования стандартных средств в медицине и здравоохранении.