***Тема №5. Основы анализа медицинских данных.***

После завершения этапа сбора данных любого научного исследования производится статистический анализ полученных результатов. К простейшим методам характеристики полученных данных относится расчет показателей описательной статистики. Эти методы позволяют охарактеризовать основные свойства изучаемых объектов, т.е. описать их.

**Основные понятия описательной статистики**

**Вариационный ряд** – ряд числовых измерений какого-либо признака, отличающихся друг от друга по своей величине и расположенных в определенном порядке (возрастания или убывания).

Каждое числовое значение в вариационном ряду называют **вариантой** ().При большой численности наблюдений некоторые варианты повторяются. В связи с этим в вариационном ряду принято выделять частоты (р). **Частота данной варианты** *–* это количество элементов совокупности, имеющих одинаковое числовое значение*.*

**Виды вариационных рядов:**

1. В зависимости от вида случайной величины выделяют два следующих вида:

- дискретный ряд;

- непрерывный ряд.

2. В зависимости от группировки вариант:

- несгруппированный;

- сгруппированный (интервальный):

3. В зависимости от частоты, с которой каждая варианта встречается в вариационном ряду:

- простой (р =1);

- взвешенный (р>1).

Вариационный ряд можно разбивать на отдельные (по возможности равные) части, которые называются квантилями. Наиболее часто употребляемые квантили представлены в таблице:



**Средняя величина** – это обобщающий показатель статистической совокупности, который погашает индивидуальные различия значений статистических величин, позволяя сравнивать разные совокупности между собой.

В зависимости от характера задачи пользуются тем или иным видом средней величины. К ним принадлежат среднее арифметическое, мода, медиана, степенные средние (среднее гармоническое, среднее геометрическое и т.п.).

Пусть имеется n объектов, для которых измерена некоторая характеристика, и получены значения , , ..., . **Среднее арифметическое** этих n значений обозначают через М (иногда также обозначается ) и определяют как

это также может быть записано следующим образом:

Среднее арифметическое может быть вычислено только для количественных величин.

**Медиана**, или средняя точка, может быть вычислена как для порядковых, так и для количественных данных. Если все элементы совокупности размещены в порядке возрастания или убывания числовых значений признака, то медиана – это такое значение признака, которое делит всю совокупность пополам.

Итак, количество элементов совокупности, имеющих значение признака, меньшее медианы, равно количеству элементов со значением признака, большим медианы. Будем обозначать медиану символом Ме.

При нахождении медианы дискретного вариационного ряда следует различать два случая:

1) объем совокупности нечетный;

2) объем совокупности четный.

Если объем совокупности нечетный и равен (2n+1), и варианты размещены в порядке возрастания их значений:

то .

Если же количество элементов четное и равно 2n, то нет варианты, которая бы делила совокупность на две равные по объему части:

поэтому в качестве медианы условно берется полусумма вариант, находящихся в середине вариационного ряда:

Медиана обладает важными свойствами, которые в некоторых случаях дают ей преимущество перед другими средними величинами. Например, если при упорядоченном размещении некоторого признака "крайние" значения сомнительные и к тому же резко отличаются от основной массы данных, то в качестве меры центральной тенденции целесообразно использовать медиану. Это связано с тем, что на ее величину эти "крайние" значения никакого влияния не оказывают, а в то же время они могут существенным образом повлиять на значение среднего арифметического.

Среднее арифметическое является хорошей мерой центральной тенденции для количественных данных, не имеющих выбросов; медиана - для порядковых данных и для количественных данных, в том числе и при наличии выбросов. Подобная характеристика нужна и для номинальных данных. Такой характеристикой является мода. Она может быть применена как для неупорядоченных категорий, так и для упорядоченных, а также для количественных данных.

**Мода** – это такое значение признака, которое встречается наиболее часто. В случае дискретных рядов вычислить моду нетрудно. Достаточно найти варианту, которая имеет наибольшую частоту или относительную частоту, это и будет мода. Будем обозначать моду символом Мо.

Если все значения в вариационном ряде встречаются одинаково часто, то считают, что этот ряд не имеет моды.

Если два соседних значения вариационного ряда имеют одинаковую частоту, и она больше частоты любого другого значения, то считают, что мода равняется среднему арифметическому этих значений.

Если два несоседних значения вариационного ряда имеют одинаковую частоту, и она больше частоты любого другого значения, то считают, что вариационный ряд имеет две моды, а соответствующее распределение называют бимодальным.

*Пример использования моды в медицинских исследованиях:*

*Требуется определить среднюю длительность госпитализации рабочих промышленных предприятий в связи с производственным травматизмом.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Число дней госпитализации* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *Итого* |
| *Число рабочих* | *6* | *18* | *14* | *10* | *6* | *3* | *2* | *1* | *60* |

*При визуальном анализе графического изображения видно, что ряд распределения несимметричен: вершина распределения сдвинута в начало ряда. Если определять среднюю величину на основе среднего арифметического (М), то средняя длительность одной госпитализации составит 4,2 дня. Однако, чаще всего (Мо) длительность госпитализации составляла 3 дня.*

*Для правильного выбора пути статистического анализа необходимо знать вид распределения изучаемого признака.*

*Под видом распределения случайной величины понимают соответствие, устанавливаемое между всеми возможными числовыми значениями случайной величины и вероятностями их появления в совокупности. Вид (закон) распределения может быть представлен:*

*- аналитической зависимостью в виде формулы;*

*- в виде графического изображения;*

*- в виде таблицы.*

**Виды распределений**

Если выйти на улицу любого города и случайным образом выбранных прохожих спросить о том, какой у них рост, вес, возраст, доход и т.п., а потом построить график любой из этих величин, то получится функция распределения данной величины. В зависимости от исследуемого признака получаемые графики могут быть различны.

Посмотрим, как можно построить такой график на примере данных роста. Сначала, просто запишем результаты исследования. Потом, отсортируем всех людей по группам, так чтобы каждый попал в свой диапазон роста, например, "от 180 до 181 включительно". После этого необходимо посчитать количество людей в каждой подгруппе (диапазоне) – это будет частота попадания роста жителей города в данный диапазон. Если затем эти частоты построить по оси у, а диапазоны отложить по оси х, можно получить гистограмму – упорядоченный набор столбиков, ширина которых равна, в данном случае, одному сантиметру, а высота будет равна той частоте, которая соответствует каждому диапазону роста. Если зарегистрированных данных было достаточно много, то полученный график будет выглядеть примерно так:



Дальше можно уточнить задачу. Каждый диапазон разбить на десять, жителей рассортировать по росту с точностью до миллиметра. Диаграмма станет глаже, но уменьшится по высоте, т.к. в каждом маленьком диапазоне количество жителей уменьшается. Если гипотетически повторить эту процедуру несколько раз, будет вырисовываться колоколообразная фигура, которая характерна для нормального (или Гауссова) распределения:



Стандартизированные кривые нормального распределения, значения функций которых приводятся в таблицах книг по статистике, всегда имеют суммарную площадь под кривой равную единице. Это связано с тем, что вероятность достоверного события всегда равна 100% (или единице), а для любого человека иметь хоть какое-то значение роста – достоверное событие.

Выделяют большое количество видов распределения признака в статистической совокупности. Остановимся на их краткой характеристике:

* нормальное распределение
* асимметричное распределение
* правостороннее
* левостороннее
* бимодальное
* альтернативное распределение

**Нормальное** (Гауссово, симметричное, колоколообразное) **распределение** – одно из самых важных распределений в статистике. Оно характеризуется тем, что наибольшее число наблюдений имеет значение, близкое к среднему, и чем больше значения отличаются от среднего, тем меньше таких наблюдений. Примерами характеристик, подчиняющихся нормальному распределению, являются показатели роста, веса, какие-либо биохимические показатели крови.

Гауссово распределение характеризует распределение непрерывных случайных величин и встречается в природе наиболее часто, за что и получило название «нормального».

Кривая нормального распределения имеет следующие свойства:

* колоколообразна (унимодальна);
* симметрична относительно среднего;
* сдвигается вправо, если среднее увеличивается, и влево, если среднее уменьшается (при постоянной дисперсии).

Среднее арифметическое, мода и медиана при нормальном распределении равны и соответствуют вершине распределения:



Нормальное распределение описывает явления, которые носят вероятностный, случайный характер, а также совместное воздействие на изучаемое явление небольшого числа случайно сочетающихся факторов. Однако, если какой-либо фактор играет преобладающую роль, то распределение не будет подчиняться Гауссову закону. Например, при исследовании показателя сахара крови для больных сахарным диабетом кривая распределения, имеющая симметричную форму для совокупности здоровых пациентов, станет несимметричной, и ее максимум сместится вправо (левостороннее асимметричное распределение).

При асимметричном распределении данных наиболее полезной мерой центральной тенденции становится медиана. Это связано с тем, что на простую среднюю арифметическую сильно влияют экстремальные (очень высокие или очень низкие) значения, из-за чего она может стать причиной неверной интерпретации результатов. Медиана же менее подвержена влиянию экстремальных величин.

Если график распределения имеет правостороннюю асимметрию ("хвост" вправо, в вариационном ряду преобладают варианты меньших значений), то в этом случае мода размещена левее, а среднее арифметическое (на рисунке обозначено как ) – правее медианы:



Обратное расположение имеет место при левосторонней асимметрии графика. При этом, чем больше асимметричен график, тем больше расстояние между его средними точками.

Проиллюстрируем важность выбора медианы, а не среднего арифметического значения на следующем примере. График заработной платы для жителей России имеет правостороннее асимметричное распределение (большинство людей имеет небольшую заработную плату). В силу того, что разброс минимальной и максимальной величин заработной платы очень велик, экстремальные значения сильно сказываются на значении среднего арифметического М (на рисунке обозначено как ). В связи с этим М сильно сдвигается в сторону «хвоста» распределения (вправо) и не может характеризовать заработную плату, соответствующую большей части населения.

Бимодальное (двугорбое) распределение наблюдается тогда, когда исследуемый признак анализируется вне однородной совокупности и, следовательно, необходимо учитывать два средних значения признака для достоверного анализа. Пример: при оценке физического развития детей подростков распределение роста будет иметь две моды (соответствующие девочкам и мальчикам).

Альтернативное распределение наблюдается в том случае, когда значения исследуемого признака распределяются по принципу: «да/нет», т.е. взаимоисключают друг друга. Подобное распределение характерно для описания качественных признаков (пример: мужской, женский пол).

Использование средних величин в медицине и здравоохранении:

а) для оценки состояния здоровья — например, параметров физического развития (средний рост, средний вес, средний объем жизненной емкости легких и др.), соматических показателей (средний уровень сахара в крови, средний пульс, средняя СОЭ и др.);

б) для оценки организации работы лечебно-профилактических и санитарно-противоэпидемических учреждений, а также деятельности отдельных врачей и других медицинских работников (средняя длительность пребывания больного на койке, среднее число посещений за 1 ч. приема в поликлинике и др.);

в) для оценки состояния окружающей среды.

В медицинских исследованиях из средних величин наиболее часто используется среднее арифметическое. В то же время, у больных людей значения многих физиологических параметров имеют асимметричное распределение, ввиду того, что изменяются в сторону увеличения или уменьшения под влиянием заболевания. Поэтому для характеристики центральной тенденции их распределения, во многих случаях, более обоснованным является как раз использование медианы, а не средней арифметической.