***Тема №6. Анализ распределения данных.***

**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 $\overbar{x}$) – 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 $\overbar{x}$). 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.