ТЕМА №6
ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

ПРОГРАММНЫЕ ВОПРОСЫ:

6.1.Задачи математической статистики. Генеральная и выборочная совокупности, способы отбора, представительность выборки.
6.2. Статистическое распределение выборки. Дискретный и интервальный ряды распределения.
6.3. Графическое представление статистических распределений выборок.
6.4. Эмпирическая функция распределения.
6.5. Понятие о несмещенности, состоятельности и эффективности оценок параметров распределения.
6.6. Выборочная средняя, выборочная и исправленная дисперсии.
6.7. Доверительный интервал для оценки математического ожидания нормального распределения. Распределение Стьюдента.
6.8. Понятие нормы для медицинских показателей.

6.1.Задачи математической статистики. Генеральная и выборочная совокупности, способы отбора, представительность выборки

Математическая статистика **–** раздел прикладной математики, непосредственно примыкающий к теории вероятностей. Основное отличие математической статистики от теории вероятностей состоит в том, что в математической статистике рассматриваются не действия над законами распределения и числовыми характеристиками случайных величин, а приближенные методы отыскания этих законов и характеристик по результатам экспериментов.

Разработка методов получения, описания и анализа экспериментальных данных, определенных в результате исследования случайных явлений, составляет предмет специальной науки – математической статистики. Эти данные принято называть ***статистическими***. Статистические данные часто можно рассматривать как совокупность экспериментальных результатов, которые представляют собой набор возможных значений случайных однородных величин (роста, массы тела, содержания сахара в крови, длительности пребывания больного на койке и т. д.).

Фундаментальными понятиями математической статистики являются *генеральная совокупность* и *выборочная совокупность (выборка)*.

Генеральная совокупность **–** это множество подлежащих статистическому изучению однородных объектов, которые характеризуются качественными или количественными признаками. Например: все жители Беларуси в фиксированный момент времени, или только все мужчины, или женщины, или дети; множество действительных чисел, лежащих между 0 и 1; количество больных ревматизмом на земном шаре и т.д. Число объектов генеральной совокупности называют ее ***объемом*** и обозначают ***N***.

Чтобы изучить генеральную совокупность по какому-либо из её количественных признаков X (острота зрения, показатели анализа крови и т.д.), нужно определить закон распределения данного признака и основные характеристики этого распределения (математическое ожидание, дисперсию). Однако на практике это сложно сделать (либо физически невозможно, либо экономически невыгодно). Поэтому исследуют только часть объектов, так называемую *выборку***.**

Выборочная совокупность – множество объектов, случайно отобранных из интересующей нас генеральной совокупности для конкретного статистического исследования. Число объектов выборки называют ее ***объемом*** и обозначают ***n***. Например, для контроля качества растворов в ампулах для инъекций на отсутствие в них механических загрязнений из серии 5000 ампул отбирают 150 ампул (N=5000 – объем генеральной совокупности, n=150 – объем выборки).

Исследование выборок дает приближенное оценочное значение для интересующего нас параметра. Следовательно, постоянная величина – значение нужной характеристики для генеральной совокупности – заменяется значением случайной величины, полученным по результатам выборки на основании некоторого правила. ***Главная цель выборочного метода*** – по вычисленной характеристике выборки как можно точнее определить соответствующую характеристику генеральной совокупности. Это возможно лишь в том случае, когда отобранная для работы часть объектов *репрезентативна* целому, т.е. типична, обладает теми же основными чертами, что и все целое. Иначе говоря, выборка должна быть *представительной*, т.е. по возможности полнее «представлять» свою генеральную совокупность. Это одно из важнейших требований, предъявляемых к выборке, невыполнение которого ведет к грубым ошибкам и обесценивает результаты исследования. Например, если при изучении заболеваемости населения республики (генеральная совокупность) ишемической болезнью сердца в качестве выборки будет взята группа студентов, то результаты окажутся ошибочными, поскольку свойства выборки не будут соответствовать свойствам генеральной совокупности, как и в случае, когда в качестве выборки будут взяты только пациенты кардиологического

диспансера. Репрезентативность выборки обеспечивается ее достаточным объемом и определенными правилами ее формирования.

В зависимости от техники отбора объектов из генеральной совокупности выборки делятся на *повторные* и *бесповторные*.

Если выборку отбирают по одному объекту, который обследуют и снова возвращают в генеральную совокупность, то выборка называется ***повторной***. Если объекты выборки не возвращаются в генеральную совокупность, то выборка называется ***бесповторной***. На практике обычно пользуются бесповторной выборкой.

На практике применяются различные способы отбора. Различают ***случайный отбор***, т. е. проводимый с помощью какого-либо случайного механизма, и ***неслучайный*** (по закономерности). В статистике применяется в основном случайный отбор как более надежный в отражении свойств генеральной совокупности.

Простым случайным отбором называется отбор, удовлетворяющий следующим требованиям:

1. выбор является случайным;
2. каждый элемент совокупности может быть выбран;
3. каждый элемент выбирается независимо от остальных;
4. все элементы выборки получаются в равных условиях.

Осуществить простой отбор можно различными способами. Например, для извлечения n объектов из генеральной совокупности объема N поступают так: выписывают номера от *1* до N на карточках, которые тщательно перемешивают и наугад вынимают одну карточку, объект, имеющий одинаковый номер с извлеченной карточкой, подвергают обследованию; затем карточка возвращается в пачку и процесс повторяется, т. е. карточки перемешиваются, наугад вынимают одну из них и т. д. Так поступают n раз; в итоге получают ***простую случайную повторную*** выборку объема n.

Если извлеченные карточки не возвращать в пачку, то выборка будет ***простой случайной бесповторной***.

Так можно выбирать группу людей для обследования, ампулы партии для испытания, лекарственные препараты для контроля и т. д.

В реальных условиях простой случайный отбор не всегда осуществим. Он является как бы эталонным идеальным отбором. Его нельзя, например, осуществить из бесконечной генеральной совокупности (время обслуживания, отклонение результата измерения от нормы), из генеральной совокупности, образование которой не завершено и может продолжаться бесконечно долго.

Виды реальных отборов:

1. ***Типическим*** называют отбор, при котором объекты отбираются не из всей генеральной совокупности, а из каждой ее «типической» части. Например, если ампулы изготовляют на нескольких станках, то

отбор производят не из всей совокупности ампул, произведенных всеми станками, а из продукции каждого станка в отдельности. Типическим отбором пользуются тогда, когда обследуемый признак заметно колеблется в различных типических частях генеральной совокупности.

2. ***Механическим*** называют отбор, при котором элементы генеральной совокупности выбираются по какой-либо закономерности. Например, измерения производятся через равные промежутки времени; контролируется каждая десятая ампула, сходящая с конвейера; каждый пятый человек по списку и т.д.

3. ***Серийным*** называют отбор, при котором объекты отбирают из генеральной совокупности не по одному, а «сериями», которые подвергаются сплошному обследованию. Например, контролю подвергается не одна таблетка лекарства, а упаковка, не один человек из какой-либо группы, а вся группа. Серийным отбором пользуются тогда, когда обследуемый признак колеблется в различных сериях незначительно.

4. ***Субъективным*** называют отбор на основе какого-либо субъективного принципа. Например, обследуются не вся партия лекарственных препаратов, а лишь одна, наиболее подозрительная часть. Он экономит время, средства, но может привести к большим ошибкам.

5. ***Выбор с помощью случайных независимых измерений*** (температура среды, загрязненность атмосферы). Характерен для естественнонаучных исследований.

На практике часто применяется комбинированный отбор, при котором сочетаются указанные выше способы.

6.2. Статистическое распределение выборки. Дискретный и интервальный ряды распределения

Итак, мы хотим знать распределение признака X в генеральной совокупности, но реально исследуем лишь некоторую выборку из неё.

В серии экспериментов, проводимых с выборкой, величина X принимает определенные значения. Эти значения, записанные для всех элементов выборки в том порядке, в котором они были получены в опытах, представляют собой ***простой статистический ряд.*** Полученные данные и подлежат статистической обработке, статистическому анализу.

Первый шаг при обработке этого материала – наведение в нем определенного порядка, ведущего к получению статистического распределения выборки.

Статистическое распределение выборки – это составление дискретного или интервального рядов, соответственно, когда количе-

ственный признак, по которому исследуют данную выборку, является дискретной или непрерывной величиной.

Дискретный ряд распределения

Пусть из генеральной совокупности извлечена выборка объемом n. В имеющемся у нас простом статистическом ряду варианта x_1 встречается (повторяется) m_1 раз, $x_2 - m_2$ раза, … $x_к - m_к$ раз и т.д. Наблюдавшиеся значения x_i признака X называют ***вариантами***, а последовательность вариант, записанную в возрастающем порядке, ***вариационным рядом*.**

Дискретный вариационный ряд удобно представить в виде таблицы, включающей в себя:

1) различные по значению варианты x_i, расположенные в определенной, заранее выбранной последовательности (обычно в порядке возрастания);

2) m_i – частоты вариантов, т.е. числа наблюдений (повторений) варианты x_i в простом статистическом ряду;

3) $p_i^* = \frac{m_i}{n}$ – относительные частоты вариант, т.е. отношения частот m_i к объему выборки n; они являются выборочными (эмпирическими) оценками вероятностей появления значений x_i.

Каждая относительная частота указывает, какая доля общего объёма выборки приходится на данное значение вариантов x_i.

Итак, для дискретной величины X вариационный ряд – статистическое распределение выборки – имеет следующий вид:

Варианта x_i $(x_1<x_2<x_3…<x_k)$	x_1	x_2	x_3	…	x_k	*Контроль*
Частота m_i	m_1	m_2	m_3	…	m_k	$\sum_{i=1}^{k} m_i = n$
Относительная частота $p_i^* = \frac{m_i}{n}$	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\frac{m_3}{n}$	…	$\frac{m_k}{n}$	$\sum_{i=1}^{k} \frac{m_i}{n} = 1$

Напомним, что под распределением дискретной случайной величины в теории вероятностей понимается соответствие между возможными значениями случайной величины и их вероятностями; в математической статистике – соответствие между наблюдаемыми вариантами x_i и их частотами или относительными частотами.

Пример 6.1: В результате отдельных испытаний активности тетрациклина были получены следующие значения (в единицах действия

на 1 мг): 925, 940, 760, 905, 995, 965, 940, 925, 940, 905. Составить ряд распределения.

Решение:

Расположив значения активности, частоты и относительные частоты в порядке возрастания, получим дискретный ряд распределения в виде таблицы:

x_i	*760*	*905*	*925*	*940*	*965*	*995*	*Контроль*
m_i	*1*	*2*	*2*	*3*	*1*	*1*	$\sum_{i=1}^{6} m_i = 10$
$p_i^* = \frac{m_i}{n}$	*0,1*	*0,2*	*0,2*	*0,3*	*0,1*	*0,1*	$\sum_{i=1}^{6} \frac{m_i}{n} = 1$

Полезность подобного представления данных очевидна по следующей причине: мы получаем практически важный результат – возможность оценить более и менее вероятные значения признака.

Интервальный ряд распределения

Интервальный ряд распределения составляется тогда, когда количественный признак *X*, является непрерывной случайной величиной, т.е. может принимать любые значения в некотором интервале.

В этом случае статистическое распределение выборки (интервальный ряд) строится следующим образом.

Для начала область изменения признака $(x_{макс} - x_{мин})$ разбивают на несколько интервалов равной ширины. Число интервалов k, как правило, не менее 5 и не более 25 приближенно определяется следующими эмпирическими формулами:

$$k = \sqrt{n}\,, \text{ или } k \approx 1 + 3{,}33 \lg n \text{ (формула Стерджесса)}$$

где n – объем выборки.

Ширину интервалов вычисляли по следующей формуле:

$$\Delta x = h = \frac{x_{макс} - x_{мин}}{k}\,.$$

Затем находят границы интервалов:

$$x_{\min} = x_0\,,\ x_1 = x_0 + h,\quad x_2 = x_1 + h,\ \ldots,\ x_{\max} = x_k\,.$$

Поскольку некоторые варианты могут являться границей двух соседних интервалов, то, во избежание недоразумений, придерживаются следующего правила: к интервалу (a,b), относят варианты удовлетворяющему неравенству $a \le x < b$.

Затем для каждого интервала подсчитывают частоты m_i и (или) относительные частоты $p_i^* = \frac{m_i}{n}$ попадания вариант в данный интервал. Нередко используют также *плотность относительной частоты*:

$$\frac{m_i}{n\Delta x} = \frac{m_i}{nh}.$$

Данную величину можно считать выборочной (эмпирической) оценкой плотности вероятности.

Рассмотренное выборочное распределение непрерывной случайной величины X– интервальный ряд – обычно представляется в виде таблицы, имеющей, в частности, следующий вид:

Интервал	x_0–x_1	x_1–x_2	x_2–x_3	…	x_{k-1}– x_k
Частота m_i	m_1	m_2	m_3	…	m_k
Относительная частота $p_i^* = \frac{m_i}{n}$	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\frac{m_3}{n}$	…	$\frac{m_k}{n}$

Пример 6.2: Анализируемый показатель X – масса тела новорожденного. Определение массы тела 100 новорожденных показало, что минимальная масса составляет 2,7кг, максимальная – 4,4 кг. Составить ряд распределения.

Решение:

Интервал $(2{,}7-4{,}4)$ кг разбиваем на 10 равных интервалов $(k=\sqrt{100}=10)$ шириной $h=\frac{4{,}4-2{,}7}{10}=0{,}17$ кг и строим интервальный ряд:

Номер интервала	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Интервал, масса тела, кг	*2,7-2,87*	*2,87-3,04*	*3,04-3,21*	*3,21-3,38*	*3,38-3,55*	*3,55-3,72*	*3,72-3,89*	*3,89-4,06*	*4,06-4,23*	*4,23-4,4*
Частота m_i	*4*	*8*	*12*	*16*	*21*	*15*	*11*	*7*	*4*	*2*
Относительная частота $p_i^* = \frac{m_i}{n}$	*0,04*	*0,08*	*0,12*	*0,16*	*0,21*	*0,15*	*0,11*	*0,07*	*0,04*	*0,02*
Плотность относительной частоты $\frac{m_i}{nh}$	*0,235*	*0,47*	*0,7*	*0,94*	*1,235*	*0,88*	*0,65*	*0,41*	*0,235*	*0,118*

Контроль: $k = 10$, $\sum_{i=1}^{10} m_i = 4 + 8 + 12 + 16 + 21 + 15 + 11 + 7 + 4 + 2 = 100 = n$ (объем выборки),

$$\sum_{i=1}^{10} \frac{m_i}{n} = 0{,}04 + 0{,}08 + 0{,}12 + 0{,}16 + 0{,}21 + 0{,}15 + 0{,}11 + 0{,}07 + 0{,}04 + 0{,}02 = 1.$$

6.3. Графическое представление статистических распределений выборок

Для получения наглядного представления о распределении выборок строят соответствующие графики, в частности, *полигон частот* или *гистограмму распределения*.

Вариационный ряд часто изображают графически в виде полигона частот или полигона относительных частот.

Для построения полигона частот на оси абсцисс откладывают варианты x_i а на оси ординат – соответствующие им частоты m_i**.** Точки (x_i, m_i) соединяют отрезками прямых.

Полигоном частот называют ломаную линию, отрезки которой соединяют точки (x_1, m_1), (x_2, m_2), …, (x_k, m_k).

Полигоном относительных частот называют ломаную линию, отрезки которой соединяют точки $(x_1, \frac{m_1}{n})$, $(x_2, \frac{m_2}{n})$, …, $(x_k, \frac{m_k}{n})$. На рисунке 6.1 показан полигон частот.

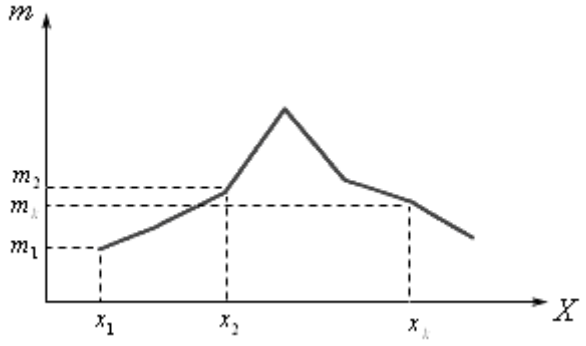

Рис.6.1

Для непрерывной случайной величины обычно строят *гистограммы частот* или *гистограммы относительных частот*.

Гистограммой частот называют диаграмму, состоящую из вертикальных прямоугольников, основаниями которых являются интервалы длиной $\Delta x = h$, а высоты равны отношению $\frac{m_i}{\Delta x}$ (плотности частоты).

Для построения гистограммы частот на оси абсцисс откладывают интервалы значений исследуемого показателя (интервалы вариант) и

на них строят прямоугольники высотой $\frac{m_i}{\Delta x}$. Площадь *i-го* прямоугольника равна $\Delta x \cdot \frac{m_i}{\Delta x} = m_i$, т.е. равна количеству вариант в *i-м* интервале. Следовательно, площадь гистограммы частот равна сумме частот для всех интервалов, иначе говоря, равна объему выборки.

Гистограмма относительных частот отличается от предыдущей гистограммы тем, что на ней высоты прямоугольников равны отношению $\frac{m_i}{n\Delta x}$, т. е. равны плотности относительной частоты (эмпирической плотности вероятности). В этом случае площадь *i-го* прямоугольника равна $\Delta x \cdot \frac{m_i}{n\Delta x} = p_i^*$ – относительной частоте вариант, попавших в *i-й* интервал (рис.6.2). Напомним, что p^* – оценка вероятности попадания значений X в выбранный интервал. Площадь гистограммы относительных частот равна сумме относительных частот для всех интервалов, т. е. равна единице.

Отметим, что гистограммой называют и фигуру, состоящую из вертикальных прямоугольников, высотами которых являются непосредственно частоты m_i для соответствующих интервалов или относительные частоты (в нормированной гистограмме), а также относительные частоты в процентах (процентная гистограмма). Два последние варианта позволяют сравнивать гистограммы, построенные на одних и тех же интервалах, но для различных выборок из той же генеральной совокупности.

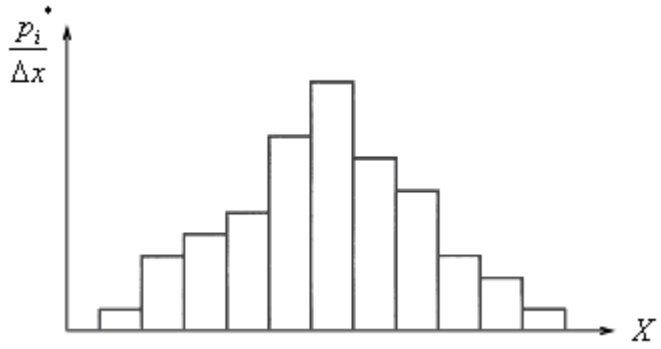

Рис. 6.2

Важно, что гистограммы можно использовать для оценки закона распределения признака в генеральной совокупности. Соединяя средние точки верхних оснований прямоугольников гистограммы относительных частот плавной линией, можно по данным выборки получить примерный вид графика зависимости плотности вероятности $f(x)$. Можно предположить, что анализируемый показатель в генеральной совокупности распределен по нормальному закону, т. е. нормальный закон является вероятностной моделью для данного признака.

Пример 6.3: Построить полигон частот и относительных частот по распределению выборки

X_i	2	3	5	6
m_i	10	15	5	20

Решение:

Полигон частот (рис. 6.3).

Полигон относительных частот (рис. 6.4).

$$P^* = \frac{m_i}{n_i}; \quad n = 10 + 15 + 5 + 25 = 50;$$

$$P_1^* = \frac{10}{50} = 0{,}2; \qquad P_2^* = \frac{15}{50} = 0{,}3; \qquad P_3^* = \frac{5}{50} = 0{,}1; \qquad P_4^* = \frac{20}{50} = 0{,}4.$$

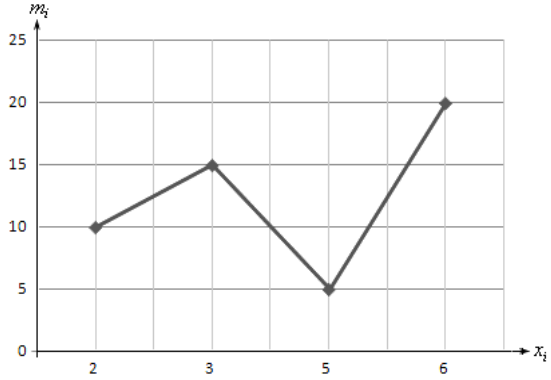

Рис. 6.3

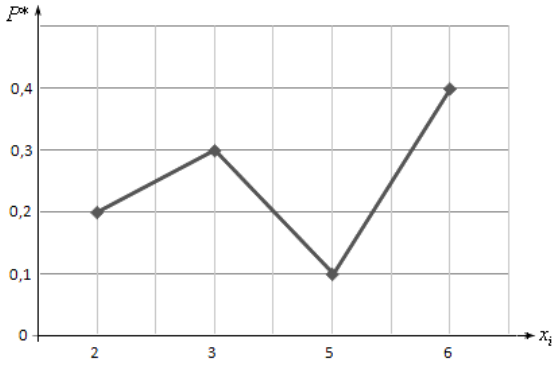

Рис. 6.4

6.4. Эмпирическая функция распределения

Предположим, что изучается некоторая случайная величина X, закон распределения которой неизвестен. Требуется определить этот закон на основании опыта или проверить экспериментально гипотезу о том, что величина X подчинена тому или иному закону. С этой целью над случайной величиной X проводится ряд независимых опытов и составляется статистическое распределение выборки количественного признака X. Чтобы получить представление о распределении случайной величины X, строят *эмпирическую функцию распределения*.

Эмпирической функцией распределения (***функцией распределения выборки***) – называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X<x$:

$$F^*(x) = \frac{m(x)}{n},$$

где $m(x)$ – число наблюдений, при которых значение признака X меньше x; n – объём выборки.

В отличии от эмпирической функции распределения $F^*(x)$ выборки, функцию распределения $F(x)$ генеральной совокупности называют ***теоретической функцией***.

Различие между эмпирической $F^*(x)$ и теоретической $F(x)$ функциями состоит в том, что $F(x)$ определяет вероятность события $X<x$, а $F^*(x)$ – относительную частоту этого же события. Поэтому эмпирическую функцию распределения выборки $F^*(x)$ можно использовать для приближённого представления теоретической функции распределения генеральной совокупности.

Функция $F^*(x)$ имеет следующие свойства:

1. Значения эмпирической функции принадлежит отрезку $[0;1]$.

2. $F^*(x)$ – неубывающая функция.

3. Если x_1 – наименьшая варианта, то $F^*(x)=0$ при $x \le x_1$; если x_k – наибольшая варианта, то $F^*(x)=1$, при $x>x_k$.

График эмпирической функции представлен на рисунке 6.5:

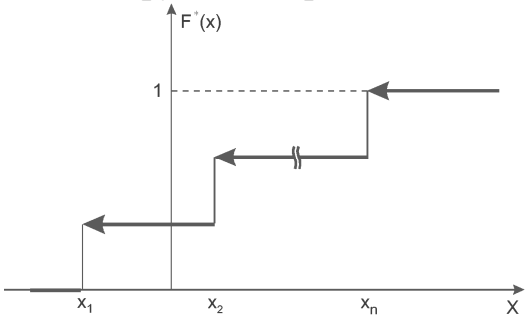

Рис. 6.5

Пример 6.4: Построить эмпирическую функцию по данному распределению выборки:

x_i	2	6	10
m_i	12	18	30

Решение:

Найдем объем выборки $n=12+18+30=60$. Наименьшая варианта равна 2, следовательно, $F^*(x)=0$ *при* $x \le 2$.

Значение $X<6$, а именно $x_1=2$, наблюдалось 12 раз, следовательно, $F^*(x)=\dfrac{12}{60}=0{,}2$ *при* $2<x \le 6$.

Значения $X<10$, а именно $x_1=2$ и $x_2=6$, наблюдалось $12+18=30$ раз, следовательно, $F^*(x)=\dfrac{30}{60}=0{,}5$ *при* $6<x \le 10$.

Так как $x = 10$ – наибольшая варианта, то $F^*(x) = 1 \; при \; x > 10$.

Искомая эмпирическая функция

$$F^*(x) = \begin{cases} 0 & при \quad x \le 2 \\ 0,2 & при \quad 2 < x \le 6 \\ 0,5 & при \quad 6 < x \le 10 \\ 1 & при \quad x > 10 \end{cases}$$

График этой функции изображен на рисунке 6.6

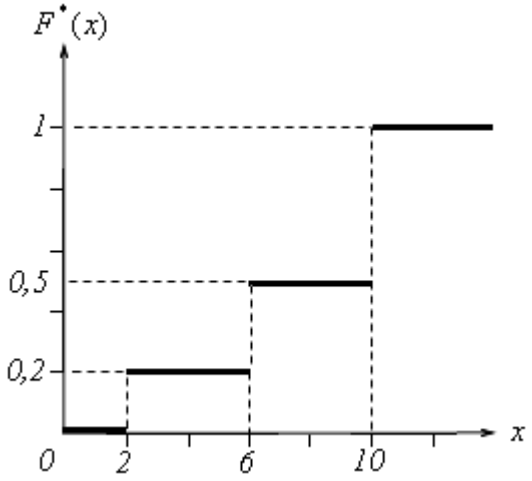

Рис 6.6

6.5. Понятие о несмещенности, состоятельности и эффективности оценок параметров распределения

К статистическому распределению выборки применимы многие характеристики распределения вероятностей. Таковы например, выборочная средняя, выборочная дисперсия, выборочное среднее квадратическое отклонение.

Характеристики статистического распределения выборки применяются для оценки неизвестных параметров теоретического распределения вероятностей. Различают *точечные* и *интервальные* оценки.

Точечной называют оценку, которая определяется одним числом. Пусть мы имеем выборку, состоящую из значений $x_1, x_2, ..., x_n$, взятую из генеральной совокупности с известным законом распределения, параметр θ имеет постоянное, но неизвестное значение. При условии, что оценке подлежит единственный параметр θ, точечная оценка представляет собой функцию от результатов наблюдений $\theta^*(x_1, x_2, ..., x_n)$.Для того, чтобы оценка давала хорошее приближение, она должна удовлетворять определённым требованиям: быть несмещённой, эффективной и состоятельной.

Точечная оценка θ^* параметра θ называется ***несмещенной***, если её математическое ожидание равно оцениваемому параметру при любом объёме выборки, т.е. $M(\theta^*) = \theta$.

Оценку, математическое ожидание которой не равно оцениваемому параметру, называют ***смещенной.***

За меру точности несмещенной оценки θ^* для параметра θ принимают дисперсию $D(\theta^*)$. Оценку с наименьшей дисперсией называют ***наилучшей.***

В качестве характеристики для сравнения точности различных оценок применяют ***эффективность*** – отношение дисперсий наилучшей оценки и данной несмещенной оценки.

При большом количестве наблюдений обычно требуется, чтобы выбранная оценка θ^* стремилась по вероятности к истинному значению неизвестного параметра θ, т.е. чтобы для любого $\varepsilon > 0$ выполнялось равенство:

$$\lim_{n\to\infty} P\left(\left|\theta^* - \theta\right| < \varepsilon\right) = 1$$

Такие оценки называют ***состоятельными.***

Из отмеченных требований, предъявляемых к оценке, наиболее важными являются требования несмещенности и состоятельности.

6.6. Выборочные характеристики распределения: выборочное среднее, выборочная дисперсия и среднее квадратическое отклонение

Методы описательной статистики – это методы описания выборок, исследуемых по количественному признаку *X*, с помощью различных числовых характеристик. Преимущество данных методов заключается в следующем: несколько простых и достаточно информативных статистических показателей, если они известны, во-первых, избавляют нас от просмотра сотен, а порой и тысяч значений вариант, во-вторых, позволяют получить более или менее точную оценку характеристик распределения признака в генеральной совокупности.

Описывающие выборку показатели разбиваются на несколько групп; в своем большинстве они имеют аналоги в виде числовых характеристик случайных величин в теории вероятностей.

Показатели положения описывают положение вариант выборки на числовой оси. Сюда относят:

а) минимальную и максимальную варианты;

б) выборочное среднее арифметическое значение (выборочное среднее).

Выборочное среднее – среднее арифметическое значение признака выборочной совокупности:

$$\overline{x_в} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \overline{x} = \frac{\sum_{i=1}^{k} x_i m_i}{n},$$

где x_i – *i-я* варианта, полученная в опыте с *i-м* элементом выборки; n – объем выборки.

Выборочная мода $Mo_в$ – варианта, которая чаще всего встречается в исследуемой выборке, т. е. имеет наибольшую частоту.

Показатели разброса описывают степень разброса данных относительно своего центра. Здесь обычно используются:

а) стандартное отклонение σ и выборочная дисперсия $D(X)$, характеризующие рассеяние вариант вокруг их среднего выборочного значения $\overline{x_в}$.

Выборочной дисперсией называют среднее арифметическое квадратов отклонения полученных значений $x_1, x_2, ..., x_n$ от выборочной средней:

$$D(X) = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_i)^2}{n}; \qquad D(X) = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_i)^2 m_i}{n}.$$

Среднее квадратическое отклонение (стандартное отклонение) рассчитывается по формуле:

$$\sigma(X) = \sqrt{D(X)}.$$

б) размах выборки – разность между максимальной и минимальной вариантами: $x_{макс} - x_{мин}$;

в) коэффициент вариации:

$$\nu = \frac{\sigma}{\overline{x_в}} \cdot 100\%.$$

Данный коэффициент применяется для сравнения величин рассеяния двух вариационных рядов. Тот из рядов имеет большее рассеяние, у которого коэффициент вариации больше.

К показателям, описывающим закон распределения**,** прежде всего относят гистограмму и полигон частот.

6.7. Доверительный интервал для оценки математического ожидания нормального распределения. Распределение Стьюдента

Под ***интервальной оценкой параметров генеральной совокупности*** понимают определение некоторого интервала, в который с заданной вероятностью попадает истинное значение исследуемого признака. Такой интервал называется ***доверительным интервалом*****,** а ве-

роятность того, что истинное значение оцениваемой величины находится внутри этого интервала, – ***доверительной вероятностью*** или ***надежностью.***

В медицинской литературе для этой величины используется термин «вероятность безошибочного прогноза». Обозначим ее γ. Значения γ задаются заранее (обычно в медико-биологических исследованиях выбирают значения γ = 0,95 = 95% или γ = 0,99 = 99%), после чего находят соответствующий доверительный интервал.[1]

Для построения надежных интервальных оценок необходимо знать закон, по которому оцениваемый случайный признак распределен в генеральной совокупности.

Рассмотрим, вначале для малых выборок ($n < 30$), как строится интервальная оценка генеральной средней $\overline{x_г} = M(X)$ признака, который в генеральной совокупности распределен по нормальному закону. В этом случае интервальной оценкой (с доверительной вероятностью γ) генеральной средней (математического ожидания) $\overline{x_г} = M(X)$ количественного признака X по выборочной средней $\overline{x_в}$ при неизвестном $\sigma_г$ является доверительный интервал:

$$\overline{x_в} - \delta < M(X) < \overline{x_в} + \delta ,$$

или в другой форме записи:

$$\overline{x_г} = M(X) = \overline{x_в} \pm \delta ,$$

где $\delta = t_{\gamma,n} \cdot \frac{S}{\sqrt{n}}$ – полуширина доверительного интервала (точность оценки); n – объем выборки; S – выборочное среднее квадратическое отклонение; $\frac{S}{\sqrt{n}}$ – стандартная ошибка выборочного среднего; $t_{\gamma,n}$ – коэффициент Стьюдента (его значения либо определяются по соответствующим таблицам, либо содержатся в программных статистических пакетах обработки данных).

Анализ формулы $\overline{x_в} - \delta < M(X) < \overline{x_в} + \delta$ показывает, что:

а) чем больше доверительная вероятность γ, тем больше коэффициент $t_{\gamma,n}$ и шире доверительный интервал;

б) чем больше объем выборки n, тем уже доверительный интервал.

При большой выборке (n>30) полуширину доверительного интервала δ определяют по соотношениям:

$$\delta = 1{,}96 \cdot \frac{S}{\sqrt{n}} \text{ – при } \gamma = 95\% \text{ или } \delta = 2{,}58 \cdot \frac{S}{\sqrt{n}} \text{ при } \gamma = 99\%.$$

---

[1] Иногда вместо доверительной вероятности используется величина α = 1 - γ, которая называется уровнем значимости.

Подобные интервальные оценки с заданной надежностью даются и тогда, когда рассматриваемый случайный признак распределен в генеральной совокупности не по нормальному, а по другим законам.

Распределение Стьюдента

Пусть случайная величина X генеральной совокупности распределена нормально, причем среднее квадратическое отклонение σ неизвестно. Требуется оценить неизвестное математическое ожидание μ при помощи доверительных интервалов.

По данным выборки из независимых наблюдений можно получить случайную величину

$$T = \frac{(\bar{x} - M(X))}{S_{\bar{x}}},$$

которая имеет распределение Стьюдента с $f = n - 1$ степенями свободы; здесь $\bar{x}$ – выборочная средняя, $S_{\bar{x}}$ – оценка среднего квадратического отклонения выборочной средней, $M(X)$ – математическое ожидание.

Плотность вероятности величины T (ее возможные значения обозначены через t) выражается формулой

$$\varphi(t) = \frac{\left(\frac{f+1}{2}\right)}{\sqrt{\pi f}\,(f/2)!\,(1 + t^2/f)^{\frac{f+1}{2}}}$$

Мы видим, что распределение Стьюдента определяется параметром n – объемом выборки (или, что то же, числом степеней свободы $f = n - 1$) и не зависит от неизвестных параметров $M(X)$ и σ; эта особенность является его большим достоинством).

Кривая плотности вероятности приведена на рисунке 6.7. Как видно из рисунка, кривая распределения симметрична относительно оси, проходящей через $t = 0$; ее ветви асимптотически приближаются к оси Ot. С ростом числа степеней свободы распределение Стьюдента приближается к нормальному и уже при $n \geq 30$ практически не отличается от него. Следовательно, при оценке неизвестных параметров по выборке малого объема $n < 30$ пользуются распределением Стьюдента.

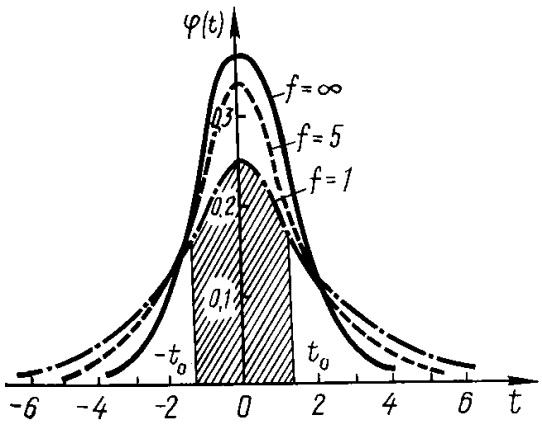

Рис. 6.7

Если известна функция плотности вероятности, можно найти вероятность того, что в результате испытания будет получено значение *t*, не превосходящее заданного t_0 по абсолютной величине: $|t| < t_0$, т.е. $\gamma = P(-t_0 < t < t_0)$. На рис. 6.7 эта вероятность для $f = 1$ численно равна площади заштрихованной фигуры и может быть вычислена по формуле

$$P(-t_0 < t < t_0) = \int_{-t_0}^{t_0} \varphi(t)dt .$$

Наоборот, для заданных *n* и γ можно указать такое t_0 , что случайно выбранное *t* должно находиться в пределах от – t_0 до t_0, т.е. с вероятностью γ выполняется неравенство

$$\frac{\left|\bar{x} - M(X)\right|}{S_{\bar{x}}} = |t| < t_0 ,$$

где коэффициент Стьюдента t_0 зависит от числа степеней свободы $f = n - 1$ и доверительной вероятности γ. Поэтому его обозначают $t_{\gamma,f}$.

Из последнего неравенства следует:

$$\left|\bar{x} - M(X)\right| \le t_{\gamma,f} \cdot S_{\bar{x}},$$

откуда $\bar{x} - t_{\gamma,f} \cdot S_{\bar{x}} \le M(X) \le \bar{x} + t_{\gamma,f} \cdot S_{\bar{x}}$.

Таким образом, интервальной оценкой математического ожидание является доверительный интервал

$$\left(\bar{x} - t_{\gamma,f} \cdot S_{\bar{x}};\ \bar{x} + t_{\gamma,f} \cdot S_{\bar{x}}\right)$$

6.8. Понятие нормы для медицинских показателей

«Нормальные» значения медико-биологических показателей являются своеобразным стандартом, характеризующим состояние здоровья человека.

Обычно используют два типа норм – *точечную норму* и *нормальный диапазон*, причем при их установлении работают с выборками достаточно большого объема. ***Точечную норму*** определяют по значению центра распределения. ***Нормальные диапазоны*** в большинстве случаев устанавливаются так, чтобы внутрь их границ гарантированно попадали *95 %* случайно отобранных здоровых людей. Когда соответствующий показатель – случайная величина – распределен по нормальному закону, точечной нормой для него считается $\overline{x_в} = \overline{x_г}$, а нормальный диапазон определяется так: $\overline{x_в} \pm 1{,}96S = \overline{x_г} \pm 1{,}96\sigma_г$; иногда используют менее точное приближение, заменяя *1,96* на *2*.

Очень часто нормальные значения некоторого показателя неодинаковы у лиц, живущих в разных географических регионах, у мужчин и женщин; у лиц разных возрастных групп. Поэтому при установле-

нии нормального значения необходимо указывать популяционные группы, к которым оно относится.