ПРАКТИЧЕСКОЕ ЗАНЯТИЕ № 3

ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ

Цель работы

1. Изучить основы первичной обработки данных;
2. Научиться строить гистограмму и полигон частот;
3. Научиться вычислять выборочное среднее значение, дисперсию и среднее квадратическое отклонение.

Математическая статистика – это раздел математики, изучающий приближенные методы отыскания законов распределения и числовых характеристик по результатам эксперимента.

В математической статистике принято выделять два основных направления исследований:

1. Оценка параметров генеральной совокупности.
2. Проверка статистических гипотез (некоторых априорных предположений).

В этом практическом занятии познакомимся как сжато описать выборочные дпнные.

Основными понятиями математической статистики являются: **генеральная совокупность, выборка, теоретическая функция распределения**.

Генеральная совокупность – это множество всех мыслимых значений наблюдений (объектов), однородных относительно некоторого признака, которые могли быть сделаны. Число всех наблюдений, составляющих генеральную совокупность, называется ее объемом N. Например, популяция представляет собой множество индивидуумов. Изучение целой популяции трудоемко и дорого и, может быть, просто невозможно. Поэтому собирают данные по выборке индивидуумов, которых считают представителями этой популяции, позволяющими сделать вывод относительно этой популяции.

Выборка – это совокупность случайно отобранных наблюдений (объектов) для непосредственного изучения из генеральной совокупности. Объем выборки *n*. Выборка обязательно должна удовлетворять условию **репрезентативности**, т.е. давать обоснованное представление о генеральной совокупности. Как сформировать репрезентативную (представительную) выборку? В идеале стремятся получить случайную (**рандомизированную)** выборку. Для этого составляют список всех индивидуумов в популяции и случайно их отбирают. Но иной раз затраты при составлении списка могут оказаться недопустимыми и тогда берут приемлемую выборку, например,

одну клинику, больницу и исследуют всех пациентов в этой клинике с данным заболеванием.

Каждый элемент выборки x_i называется **вариантой**. Число наблюдений варианты n_i называется **частотой встречаемости**. Последовательность вариант, записанных в **возрастающем порядке**, называется **вариационным рядом.**

Статистическое распределение **– это совокупность вариант x_i и соответствующих им частот n_i.**

Пример 3.1. Задано распределение частот выборки объема 20.

x_i	2	6	12
n_i	3	10	7

Написать распределение относительных частот.

Р е ш е н и е . Найдем относительные частоты. Для этого разделим частоты на объем выборки:

$$\frac{n_1}{n}=\frac{3}{20}=0{,}15;\qquad \frac{n_2}{n}=\frac{10}{20}=0{,}5;\qquad \frac{n_3}{n}=\frac{7}{20}=0{,}35.$$

Распределение относительных частот имеет вид:

x_i	2	6	12
ω_i	0,15	0,5	0,35

Контроль: 0.15+0.5+0.35=1.

Для наглядного представления статистического распределения пользуются графическим изображением вариационных рядов: полигоном и гистограммой.

Гистограмма частот – это ступенчатая фигура, состоящая из смежных прямоугольников, построенных на одной прямой, основания которых одинаковы и равны ширине класса, а высота равна или частоте попадания в интервал n_i или относительной частоте $\frac{n_i}{n}$.

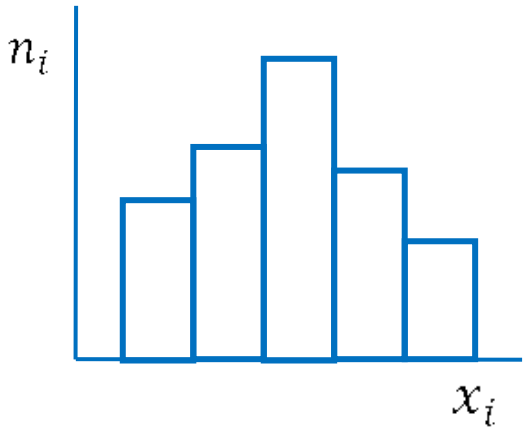
Ширину интервала i можно определить по формуле Стерджеса:

$$i=\frac{x_{\max}-x_{\min}}{1+3.32\lg n},$$

где $x_{\max}$ – максимальное,

а $x_{\min}$ – минимальное значения вариант, а их разность носит название вариационный размах.

n – объем выборки.



Полигон частот – ломаная линия, отрезки которой соединяют точки с координатами $x_i;n_i$.

Пример 3.2. Построить дискретный вариационный ряд и начертить полигон распределения 45 абитуриентов по числу баллов, полученных ими на приемных экзаменах:

39 41 40 42 41 40 42 44 40 43 42 41 43 39 42 41 42 39 41 37 43 41

38 43 42 41 40 41 38 44 40 39 41 40 42 40 41 42 40 43 38 39 41 41 42.

Р е ш е н и е . Для построения вариационного ряда различные значения признака x располагаем в порядке их возрастания и под каждым из этих значений записываем его частоту.

x_i	37	38	39	40	41	42	43	44
n_i	1	3	5	8	12	9	5	27

Построим полигон этого распределения:

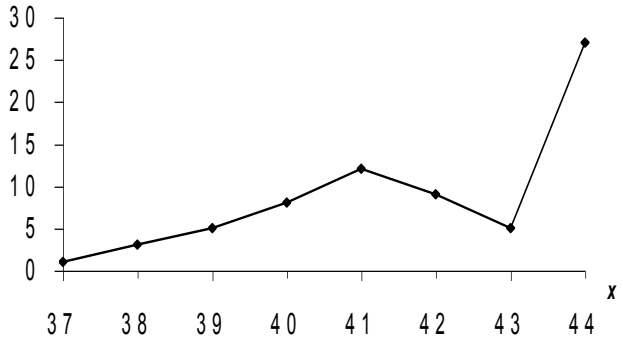


Рис.3.1. Полигон частот

Пример 3.3. Наблюдения за числом частиц, попавших в счетчик Гейгера, в течение минуты дали следующие результаты:
21 30 39 31 42 34 36 30 28 30 33 24 31 40 31 33 31 27 31 45 31 34 27 30 48 30 28 30 33 46 43 30 33 28 31 27 31 36 51 34 31 36 34 37 28 30 39 31 42 37.

Построить по этим данным интервальный вариационный ряд с равными интервалами (I интервал 20-24; II интервал 24-28 и т.д.) и начертить гистограмму.

Р е ш е н и е .

Интервал	20-24	24-28	28-32	32-36	36-40	40-44	44-48	48-52
Частота	1	4	22	8	7	4	2	2

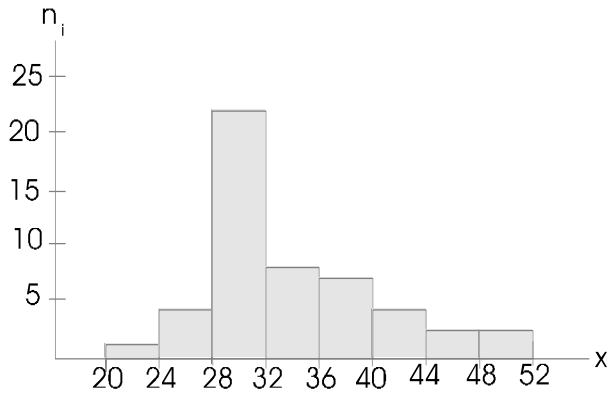


Рис.3.2. Гистограмма распределения

Статистические оценки параметров распределения. Выборочные характеристики

Характеристики положения

Мода **(M_0)** – это такое значение варианты, что предшествующее и следующее за ним значения имеют меньшие частоты встречаемости.

Для одномодальных распределений мода - это наиболее часто встречающаяся варианта в данной совокупности.

Например, мода распределения:

x_i	16	17	18	20
n_i	5	1	20	6

равна 18.

Для определения моды интервальных рядов служит формула:

$$M_o = x_{\text{ниж}} + i \cdot \left(\frac{n_2 - n_1}{2n_2 - n_1 - n_3}\right), \quad \text{где}$$

$x_{\text{ниж}}$ – нижняя граница модального класса, т.е. класса с наибольшей частотой встречаемости n_2;

n_2 – частота модального класса;

n_1 – частота класса, предшествующего модальному;

n_3 – частота класса, следующего за модальным;

i – ширина классового интервала.

Пример 3.4. Определить моду ряда распределения кальция (*мг* %) в сыворотке крови обезьян.

Классы	8,6-9,3	9,4-10.1	10,2-10,9	11,0-11,7	11,8-12,5	12,6-13,3	13,4-14,1	14,2-14,9
n_i	4	7	13	23	25	17	10	2

Р е ш е н и е . Частота модального класса n_2 = 25, его нижняя граница $x_{ниж}$ = 11.8. Частота класса, предшествующего модальному, n_1 = 23; частота класса, следующего за модальным n_3 = 17; i = 0.7. Подставим эти данные в формулу, находим:

$$M_o = 11{,}8 + 0{,}7 \cdot \left(\frac{25-23}{50-23-17}\right) = 11{,}94$$

Пример 3.5. Найдите моду распределения роста 1000 взрослых женщин:

Рост, см	Число мужчин	Рост, см	Число мужчин
143-146	1	167-170	170
146-149	2	170-173	120
149-152	8	173-176	64
152-155	26	176-179	28
155-158	65	179-182	10
158-161	120	182-185	3
161-164	181	185-188	1
164-167	201		

Р е ш е н и е. $M_o = 164 + 3 \cdot \left(\frac{201-181}{2\cdot 201-181-170}\right) = 165{,}18$

Медиана M_e – это значение признака, относительно которого ряд распределения делится на 2 равные по объему части.

Например, в распределении:

12 14 16 18 20 22 24 26 28

медианой будет центральная варианта, т.е. M_e = 20, так как по обе стороны от нее отстоит по 4 варианты.

Для ряда с четным числом членов 6 8 10 12 14 16 18 20 22 24 медианой будет полусумма его центральных членов, т.е. $M_e = \frac{14+16}{2} = 15.$

Выборочная средняя – это среднее арифметическое значение вариант статического ряда

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{K} x_i n_i .$$

Характеристики рассеяния вариант вокруг своего среднего

Выборочная дисперсия – среднее арифметическое квадратов отклонения вариант от их среднего значения:

$$S_в^2 = \frac{1}{n}\sum_{i=1}^{к}(x_i - \bar{x}_в)^2 \cdot n_i.$$

Среднее квадратическое отклонение – это квадратный корень из выборочной дисперсии:

$$S_в = \sqrt{S_в^2}\ .$$

Коэффициент вариации *CV* – это отношение среднего квадратического отклонения к средней величине признака, выраженное в процентах:

$$CV = \frac{S_в}{\bar{x}_в} \cdot 100\%.$$

Коэффициент вариации – это мера относительной изменчивости случайной величины, которая позволяет сравнивать разнородные величины, например, частоту сердечных сокращений (ЧСС, уд/мин), артериальное давление (АД, мм.рт.ст.) и температуру (t°,°C) **в единых единицах – процентах.**

Вариационный размах $\Delta = x_{\max} - x_{\min}$ – это разность между наибольшим и наименьшим значениями признака.

Пример 3.6. Выборочная совокупность задана таблицей распределения

x_i	1	2	3	4
n_i	20	15	10	5

Найти выборочную дисперсию.

Р е ш е н и е .

Найдем выборочную среднюю:

$$\bar{x}_в = \frac{20 \cdot 1 + 15 \cdot 2 + 10 \cdot 3 + 5 \cdot 4}{20 + 15 + 10 + 5} = 2.$$

Найдем выборочную дисперсию:

$$S_в^2 = \frac{20(1-2)^2 + 15(2-2)^2 + 10(3-2)^2 + 5(4-2)^2}{20 + 15 + 10 + 5} = 1.$$

Пример 3.7. Сравните 2 варьирующихся признака. Один характеризуется средней $\overline{X}_1$ = 2,4 *кг* и средним квадратическим отклонением S_1 =0,58 *кг* другой – величинами $\overline{X2}$=8,3 *см* и S_2 =1,57 *см*. Какой признак варьируется сильнее?

Р е ш е н и е .

$$CV_1 = \frac{S_1}{\bar{x}_1} \cdot 100\%,$$

$$CV_1 = \frac{0.58}{2.4} \cdot 100\% = 24.2\%, \qquad CV_2 = \frac{1.57}{8.3} \cdot 100\% = 18.9\%.$$

Ответ: первый, так как $CV_1 > CV_2$.