1. Вариационный ряд (дискретный и интервальный). Характеристики вариационного ряда: характеристики положения (среднее арифметическое и его свойства), мода, медиана, показатели вариации (размах варьирования, выборочная дисперсия и ее свойства, коэффициент вариации), асимметрия и эксцесс. Эмпирическая функция распределения. Полигон и гистограмма.

Вариационный ряд (дискретный и интервальный)

Вариационный ряд — это упорядоченная последовательность вариант (значений признака) с соответствующими им частотами или относительными частотами.

- **Дискретный вариационный ряд** строится для признаков, принимающих конечное или счетное число значений. Каждому значению x_i соответствует частота n_i (сколько раз это значение встретилось в выборке).
- **Интервальный вариационный ряд** используется для непрерывных признаков или при большом объеме выборки. Данные разбиваются на интервалы (обычно равной длины h), и для каждого интервала подсчитывается количество попадающих в него значений. Ширина интервала определяется по формуле: $h = \frac{x_{max} - x_{min}}{k}$, где k — количество интервалов.

Примеры

дискретный вариационный ряд:

x_i	0	1	2	3	Сумма
$\widehat{p}_i$	21	11	3	1	36

Другой вид вариационного ряда — варианты с относительными частотами:

x_i	0	1	2	3	Сумма
$\widehat{p}_i$	$\frac{21}{36}$	$\frac{11}{36}$	$\frac{3}{36}$	$\frac{1}{36}$	1

Интервальная таблица частость будет выглядеть следующим образом:

$(x_{i-1}, x_i]$	(2,91-3,19]	(3,19-3,47]	(3,47-3,75]	(3,75-4,03]	(4,03-4,31]	(4,31-4,59]	(4,59-4,87]
n_i	3	7	14	13	10	2	1

Или интервальная таблица относительных частот:

$(x_{i-1}, x_i]$	(2,91-3,19]	(3,19-3,47]	(3,47-3,75]	(3,75-4,03]	(4,03-4,31]	(4,31-4,59]	(4,59-4,87]
$\widehat{p}_i$	0,06	0,14	0,28	0,26	0,2	0,04	0,02

■

Характеристики положения

Позволяют определить «центр» распределения:

- **Среднее арифметическое ($\bar{x}$)** — сумма произведений вариант на их частоты, деленная на объем выборки: $\bar{x} = \frac{\sum x_i n_i}{n}$.
  - **Свойства:**
    1. **Алгебраическая сумма отклонений** вариант от их среднего значения всегда равна нулю ($\sum(x_i - \bar{x})n_i = 0$).
    2. Если все варианты изменить на постоянную величину C (вычесть или прибавить), среднее изменится на ту же величину C.
    3. Если все варианты разделить на число k, среднее также уменьшится в k раз.
- **Мода (Mo)** — наиболее часто встречающееся значение в выборке. В интервальном ряду это значение в интервале с наибольшей частотой.

Определение 5.26. Выборочной модой $\widehat{Mo}$ ДИСКРЕТНОГО вариационного ряда называется варианта x_i выборки, имеющий наибольшую частоту (относительную частоту).

Для выборки из НЕПРЕРЫВНОЙ генеральной совокупности вычисления проводят для интервального вариационного ряда. Для сгруппированной выборки сначала определяют модальный интервал, т.е. интервал с наибольшей частотой $(x_{i-1}; x_i]$. Затем, используя линейную интерполяцию, определяют выборочную моду.

Определение 5.27. Мода интервального вариационного ряда определяется по формуле:

$$\widehat{Mo} = x_{Mo} + h\frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})},$$

где x_{Mo} - начало модального интервала, $n_{Mo-1}, n_{Mo}, n_{Mo+1}$ - частоты предмодального, модального и послемодального интервалов, h - ширина частичного интервала.

Для симметричного распределения в качестве моды можно взять середину модального интервала.

- **Медиана (Me)** — значение признака, которое делит ранжированную выборку на две равные части: у половины объектов значение признака меньше медианы, у другой половины — больше.

Определение 5.28. *Выборочной медианой* ДИСКРЕТНОГО вариационного ряда назывыется число $\widehat{Me}$, которое делит вариационный ряд на две части, содержащие равное количество элементов.

Если объем выборки n - нечетное число ($n = 2m+1$), то $\widehat{Me} = x_{m+1}$, т.е. $\widehat{Me}$ является средним элементом вариационного ряда. Если же n – четное ($n = 2m$"), то $\widehat{Me} = \frac{1}{2}(x_m + x_{m+1})$.

Ясно, что для любой выборки можно подобрать бесконечно много медиан. Чтобы избежать неоднозначности, будем называть выборочной медианой число $\widehat{Me}$ такое, когда $\widehat{p}_{Me} = 0,5$, где 0,5 - ордината точки с абсциссой $\widehat{Me}$ на кривой накопленных частот.

Для того, чтобы найти медиану, нужно найти сначала медианный интервал $(x_{i-1}; x_i]$, где $\widehat{p}_{x_{i-1}} < 0,5$ и $\widehat{p}_{x_i} > 0,5$. Тогда $\widehat{Me} \in (x_{i-1}; x_i]$. Используя линейную интерполяцию, вычислить выборочную медиану.

Определение 5.29. *Выборочной медианой* ИНТЕРВАЛЬНОГО вариационного ряда называют такое число x_{Me}, когда 50% вариант выборки меньше этого значения, а 50% больше его, вычисляется по формуле

$$\widehat{Me} = x_{Me} + \frac{h \cdot (0,5 \cdot n - n_{x_{Me-1}})}{n_{Me}}, \tag{5.3}$$

где x_{Me} - начало (левая граница) медианного интервала; $n_{x_{Me-1}}$ - накопленная частота интервала, предшествующего медианному; n_{Me} - частота медианного интервала; n - объем выборки, h - ширина частичного интервала.

Показатели вариации

Показывают, насколько сильно данные разбросаны вокруг центра.

- **Размах варьирования (R)** — простейшая мера, разность между самым большим и самым маленьким значениями ($R = x_{max} - x_{min}$).
- **Выборочная дисперсия ($\hat{D}_B$)** — среднее арифметическое квадратов отклонений вариант от выборочного среднего.

$$\begin{aligned}\widehat{D}(x) &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{n}x_j^2 - (\overline{x})^2 = \\ &= \frac{1}{n}\sum_{i=1}^{k}n_i x_i^2 - (\overline{x})^2 = \sum_{i=1}^{k}\widehat{p}_i x_i^2 - (\overline{x})^2,\end{aligned} \tag{5.4}$$

  - **Свойства:** Дисперсия не меняется, если ко всем вариантам прибавить одно и то же число.
  - **Исправленная дисперсия (s^2):** Используется для получения несмещенной оценки генеральной дисперсии, особенно на малых выборках. Она вычисляется как $s^2 = \frac{n}{n-1}D_B$.
- **Среднее квадратическое отклонение ($\hat{\sigma} = \sqrt{\hat{D}_B}$)** — корень квадратный из дисперсии.
- **Коэффициент вариации (V):** выражает стандартное отклонение в процентах от среднего арифметического: $V = \frac{\sigma}{\bar{x}} \cdot 100\%$. Он позволяет сравнивать изменчивость разных признаков.

Асимметрия и эксцесс

Характеризуют форму распределения:

Определение 5.31. Коэффициенты *выборочной асимметрии* $\widehat{A}$ и *выборочного эксцесса* $\widehat{E}$ вычисляются по формулам:

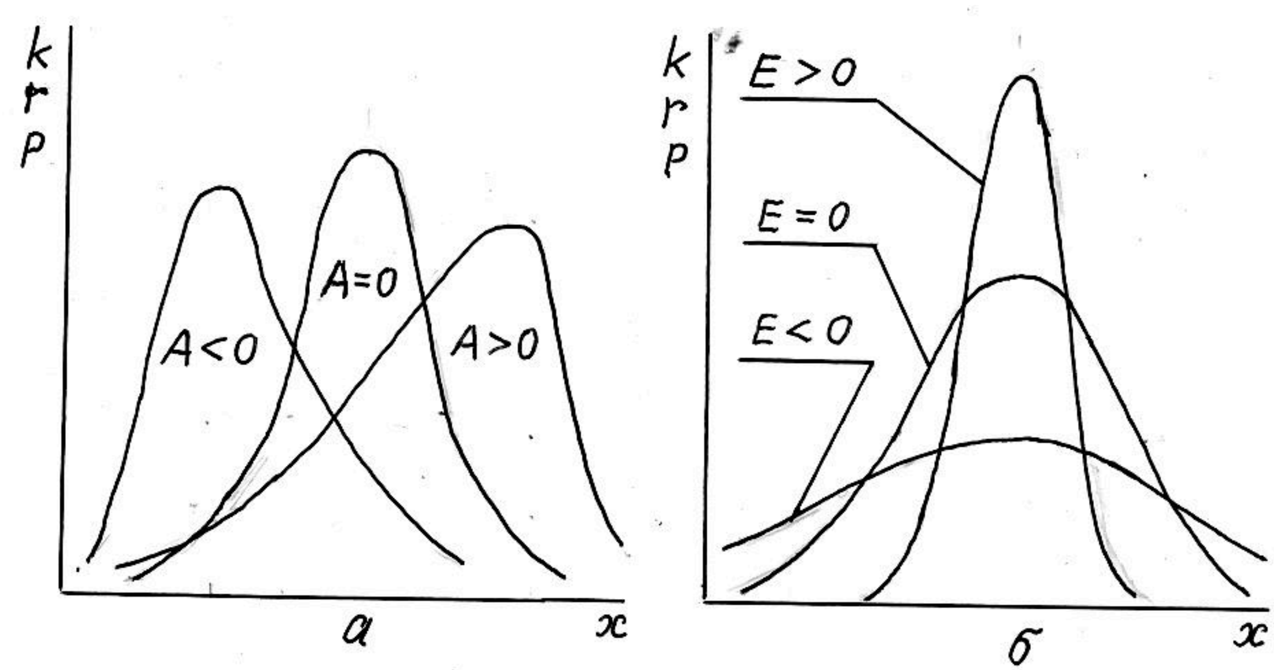
$$\widehat{A} = \frac{1}{\widehat{\sigma}^3} \sum_{i=1}^{k} (x_i - \overline{x})^3 \cdot \nu_i, \qquad \widehat{E} = \frac{1}{\widehat{\sigma}^4} \sum_{i=1}^{k} (x_i - \overline{x})^4 - 3. \tag{5.6}$$

Для симметричного распределения $\widehat{A} = 0$; если $\widehat{A} > 0$, то наблюдается правосторонняя асимметрия (справа на графике плонтости распределения более длинный «хвост»), при $\widehat{A} < 0$ имеем левостороннюю асимметрию (более длинный «хвост» распределения слева).

Коэффициент эксцесса $\widehat{E}$ характеризует крутость распределения по сравнению с нормальным. Если кривая плотности распределения более островершинная, чем нормальная, то $\widehat{E} > 0$, если более плоская, чем нормальная, то $\widehat{E} < 0$.

Отметим в заключении, что рассчитанные выборочные характеристики не являются оценками параметров генеральной совокупности. Необходимо внести определенные поправки для того, чтобы выборочные характеристики давали «хорошие» точечные оценки генеральных параметров. А для этого надо знать, какие оценки являются «хорошими».

- **Асимметрия (As)** — показатель «скошенности» графика. Если $As > 0$, распределение вытянуто вправо; если $As < 0$ — влево.
- **Эксцесс (Ek)** — показатель «островершинности». Для нормального распределения $Ek = 0$. Если $Ek > 0$, пик графика более острый, чем у нормальной кривой; если $Ek < 0$ — более плоский.



- **Эмпирическая функция распределения** $F^*(x)$ — функция, определяющая для каждого значения x относительную частоту события $X < x$. Это ступенчатая функция, которая растет от 0 до 1.

- **Полигон частот** — ломаная линия, соединяющая точки (x_i, n_i), используется для визуализации дискретных рядов.

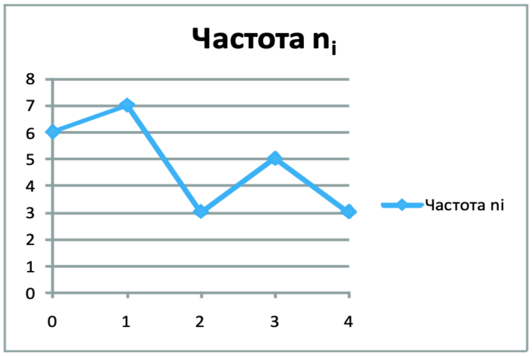


Рис.5.1. Полигон частот

- **Гистограмма** — ступенчатая фигура из прямоугольников для интервального ряда. Площадь каждого прямоугольника равна относительной частоте (или частоте) интервала.

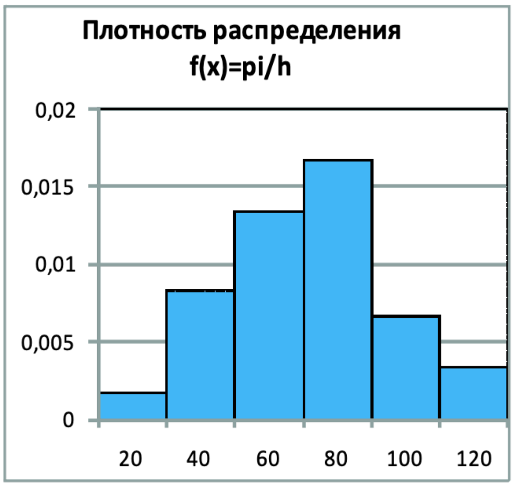


Рис.5.2. Гистограмма

Аналогия для запоминания:
Представь вариационный ряд как **результаты стрельбы по мишени**.

- **Среднее арифметическое** — это «центр тяжести» всех твоих попаданий.
- **Мода** — это то конкретное место в мишени, куда ты попадал чаще всего.
- **Дисперсия** — это твой разброс: насколько кучно лежат пули.
- **Асимметрия** покажет, есть ли у тебя систематический «занос» руки вправо или влево, а **эксцесс** — насколько предсказуемо ты попадаешь (бьешь ли ты всё время в одну точку или «сеешь» равномерно вокруг центра).