Интервальный вариационный ряд и его характеристики.

1. Построение интервального вариационного ряда.

При изучении генеральной совокупности коров по суточному удою были получены следующие данные (в порядке их поступления):

21,9	14,9	20,4	20,7	21,7	22,4	24,8	14,1	24,8	16,5
21,4	22,1	21,1	22,7	18,5	22,3	31,2	19,5	22,7	23,7
22,9	25,4	23,0	12,3	27,8	18,2	21,8	20,3	17,2	18,7
21,3	27,2	24,8	23,0	29,6	17,8	22,2	15,7	20,1	26,9
23,9	21,6	17,0	14,6	24,4	17,3	21,2	25,3	18,3	22,3
23,2	19,9	22,5	24,5	17,2	23,5	21,3	25,9	21,9	15,2
16,2	27,7	20,5	19,6	13,9	25,7	22,9	20,9	24,1	18,1
21,2	18,5	19,7	18,6	20,6	20,1	23,5	20,7	19,3	20,9
21,4	26,1	13,2	25,3	23,4	16,7	24,2	19,0	19,7	16,1
20,7	20,5	26,3	16,4	25,7	19,1	25,0	23,8	15,5	28,5

Интересующий нас признак принимает дробные, практически не повторяющиеся значения, поэтому целесообразно построить **интервальный** вариационный ряд. Для этого необходимо определить число интервалов (классов) и длину интервала (классного промежутка), после чего произвести *разноску*, т.е. подсчитать для каждого интервала число вариант, попавших в него.

Количество классов устанавливают в зависимости от степени точности, с которой ведется обработка, и количества объектов в выборке. Считается удобным при объеме выборки *(n)* в пределах от 30 до 60 вариант распределять их на 6–8 классов, при *n* от 60 до 100 вариант – на 7–8 классов, при *n* от 100 и более вариант- на 9–17 классов.

Нужное количество групп также может быть ориентировочно вычислено по формуле Стерджесса:

$$k = 1 + 3{,}332 \lg n$$

где k - число групп (классов, интервалов) ряда распределения; n – объем выборки.

Можно также использовать выражение:

$$k = \sqrt{n}.$$

При $n \leq 70$ они дают примерно одинаковые результаты.

В рассматриваемом примере по суточным удоям коров, $n = 100$. Применяя формулу Стерджесса, получим:

$$k = 1 + 3{,}332 \lg 100 = 1 + 3{,}322 \cdot 2 = 7{,}644 \approx 8.$$

Однако $\sqrt{100} = 10$. Таким образом число интервалов может быть равно 8, 9,10 и тд.(см выше).

Нахождение нужного количества групп и их размеров часто бывает взаимообусловлено. Для того, чтобы как-то определиться с числом интервалов, найдем **размах вариации** - разность между наибольшей и наименьшей вариантой:

$$R = x_{max} - x_{min}$$

где R - размах вариации, x_{min} - наименьшее знамение варьирующего признака, x_{max} - наибольшее значение варьирующего признака.

Найдем размах вариации для рассматриваемой задачи:

$$R = 31{,}2 - 12{,}3 = 18{,}9.$$

Для того чтобы найти длину интервала (величину классового промежутка) необходимо разделить размах вариации на число классов и полученную величину округлить таким образом, чтобы было удобно производить сначала разноску, а затем и различные вычисления, т.е.:

$$h \approx \frac{R}{k}$$

Предположим, что мы все-таки решили взять 10 интервалов. Тогда

$$h \approx \frac{18{,}9}{10} = 1{,}89.$$

Округлив полученную величину до целых, примем длину интервала $h = 2$.

Теперь необходимо определиться с началом первого интервала. Для этого можно использовать формулу:

$$x_1 \approx x_{\min} - \frac{h}{2}.$$

Можно за начало первого интервала принять некоторое значение, несколько меньшее $x_{\min}$, например, ближайшее целое. Так как длина интервала - целое число, это будет удобнее. Поэтому примем за начало первого интервала $x_1 = 12$.

Прибавив к началу первого интервала (нижней границе) шаг, получим верхнюю границу первого интервала и одновременно нижнюю границу второго интервала. Выполняя последовательно указанные действия, будем находить границы последующих интервалов до тех пор, пока не будет получено или перекрыто $x_{\max}$.

Таким образом, верхняя граница одного интервала одновременно является верхней границей другого интервала. Чтобы не возникало сомнений, в какой интервал отнести варианту, попавшую на границу, условимся относить ее к **верхнему** интервалу.

Замечание. Можно поступить и по-другому. Для этого необходимо установить верхние границы классов не совпадающими с последующими нижними границами, т е уменьшить их на величину, равную точности измерения признака В нашем примере это 0,1 Тогда границы классов были бы следующими- 12-13,9; 14-15,9; 16-17,9 и т д.

Составим теперь рабочую таблицу для построения **интервального** вариационного ряда и произведем подсчет частот, попавших в тот или иной интервал. Порядок ее составления и подсчета частот был детально рассмотрен в *практической работе № 1*.

Таблица 2.1.

Распределение коров по суточному удою.

№	Границы интервалов $(x_i; x_{i+1}]$	Частоты, m_i
1	12 – 14	3
2	14 – 16	6
3	16 – 18	10
4	18 – 20	15
5	20 – 22	24
6	22 – 24	19
7	24 – 26	14
8	26 – 28	6
9	28 - 30	2
10	30 – 32	1

Мы получили **интервальный вариационный ряд** - упорядоченную совокупность интервалов варьирования значений случайной величины с

соответствующими частотами попаданий в каждый из них значений величины.

После разноски частот по интервалам, для выполнения дальнейших расчетов, найдем середины интервалов, как среднее арифметическое начала и конца интервала. Кроме этого найдем и накопленные частоты. Результаты занесем в *таблицу № 2.2.*

Таблица 2.2.

№	Границы интервалов $(x_i; x_{i+1}]$	Середины интервалов $x_i^* = \frac{x_i + x_{i+1}}{2}$	Частоты, m_i	Накопленные частоты, S_i
1	12 – 14	13	3	3
2	14 – 16	15	6	9
3	16 – 18	17	10	19
4	18 – 20	19	15	34
5	20 – 22	21	24	58
6	22 – 24	23	19	77
7	24 – 26	25	14	91
8	26 – 28	27	6	97
9	28 - 30	29	2	99
10	30 – 32	31	1	100

2.Графическое изображение вариационных рядов

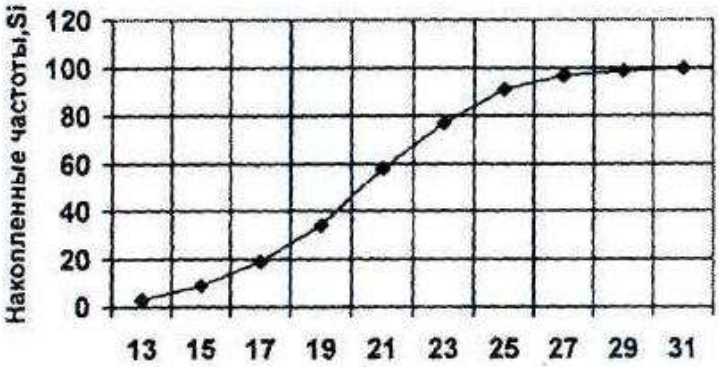
Определение 1. *Гистограммой частот (относительных частот)* называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат длины частичных интервалов, а высотами – частоты (относительные частоты).

Постоим гистограмму частот.

Замечание. Полигон частот можно построить и на графике гистограммы частот, соединив прямыми линиями середины верхних сторон прямоугольников.

Используя данные *таблицы № 2.2* построим **кумуляту** интервального вариационного ряда. Нижней границе первого интервала соответствует частота, равная нулю, верхней границе - вся частота и интервала. Верхней границе второго интервала соответствует накопленная частота первых двух интервалов (т.е. сумма частот этих интервалов) и т.д. Верхней границе последнего (максимального) интервала соответствует накопленная частота, равная сумме всех частот.

Кумулята



3. Построение функции распределения

Искомая эмпирическая функция распределения равна нулю для всех значений от $-\infty$ до 12 включительно. Действительно, пусть, например, $x = 12$.

Тогда $F_n(x)=\frac{n(12)}{100}$, где *n(12)* - число коров, удой которых меньше 12 кг. Но среди обследованных животных таких не оказалось.

Далее мы можем найти значение эмпирической функции распределения для $x=14$ (правого конца первого интервала) и не можем сделать этого ни для одной внутренней точки интервала 12-14. Пусть например $x=13$. Тогда $F_n(13)=\frac{n(13)}{100}$, где *n(13)* - число животных, удой которых меньше 13 кг. Но к

первому интервалу отнесены животные, удой которых в пределах от 12 до 14 кг. Сколько среди них животных, удой которых меньше 13 кг мы не знаем.

Замечание. Вернувшись к исходным данным, мы сможем определить это значение, но это фактически будет означать возврат к дискретному ряду, что неоправданно.

Но при $x=14$ имеем:

$$F_n(14)=\frac{3}{100}=0{,}03,$$

поскольку $n(14) = 3$ - согласно *таблице № 2.2* три особи имели удой меньше 14 кг.

В аналогичных условиях мы оказываемся при нахождении значений эмпирической функции распределения для последующих значений аргумента: мы не можем указать значения ее для всех внутренних точек каждого интервала удоя, но значения ее легко вычисляются для значений x, соответствующих правым границам каждого интервала

Например, при x=16 имеем:

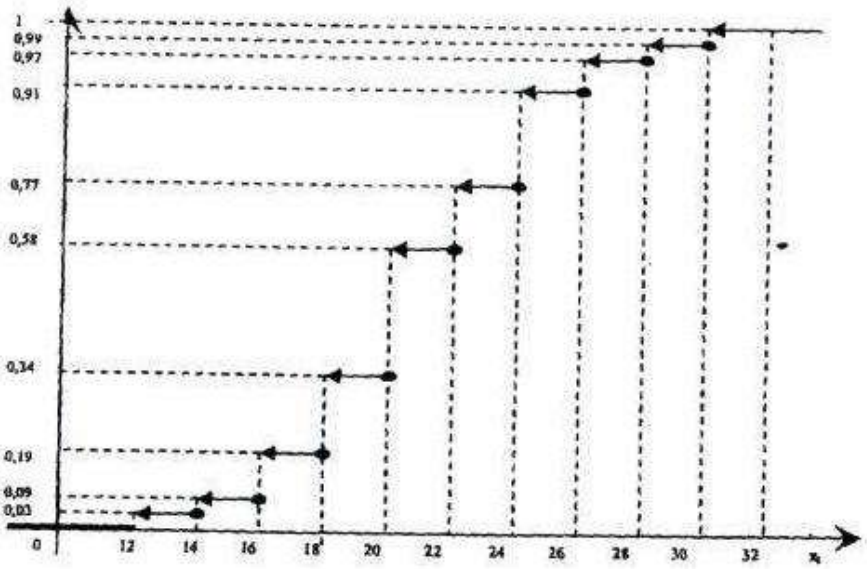
$$F_n(16) = \frac{n(16)}{100} = \frac{3+6}{100} = 0{,}09$$

так как число животных, удой которых меньше 16 находится суммированием частот интервалов 12-14 и 14-16, т.е. равно накопленной частоте интервала 14-16. Используя данные *таблицы № 2.2* по накопленным частотам, получим эмпирическую функцию распределения $F_n(x)$:

$$F_n(x) = \begin{cases} 0, & при \quad x \le 12, \\ 0{,}03 & при \quad 12 < x \le 14, \\ 0{,}09 & при \quad 14 < x \le 16, \\ 0{,}19 & при \quad 16 < x \le 18, \\ 0{,}34 & при \quad 18 < x \le 20, \\ 0{,}58 & при \quad 20 < x \le 22, \\ 0{,}77 & при \quad 22 < x \le 24, \\ 0{,}91 & при \quad 24 < x \le 26, \\ 0{,}97 & при \quad 26 < x \le 28, \\ 0{,}99 & при \quad 28 < x \le 30 \\ 1 & при \quad x > 30. \end{cases}$$

Изобразим эту функцию графически:

Эмпирическая функция распределения



4. Числовые характеристики выборки

Находим:

1. выборочную среднюю $\bar{x}$ по формуле: $\bar{x} = \frac{\sum_{i=1}^{k} x_i^* m_i}{n}$,

где x_i^* - варианта, попадающая на середину ряда; m_i - соответствующая частота, k - количество различных вариант, n - объем выборки.

5. выборочную дисперсию $D(X)$ по формуле: $D(X) = \frac{\sum (x_i^* - \bar{x})^2 m_i}{n}$.
6. среднее квадратическое отклонение σ.
7. «исправленную» дисперсию: $s^2 = \frac{n}{n-1} \cdot D(X)$.
8. «исправленное» среднее квадратическое отклонение: s.
9. коэффициент вариации: υ.

Таблица 2.3.

Вспомогательная таблица для расчета числовых характеристик ряда распределения

Интервалы $(x_i; x_{i+1}]$	Середина интервала x_i^*	Частота (m_i)	$x_i^* m_i$	$x_i^* - \bar{x}$	$(x_i^* - \bar{x})^2 m$	$\frac{x_i^* - \bar{x}}{\sigma}$	$\left(\frac{x_i^* - \bar{x}}{\sigma}\right)^3$	$\left(\frac{x_i^* - \bar{x}}{\sigma}\right)^4 m$
12 – 14	13	3	39	-8,25	204,1875	-2,285	-35,806	81,783
14 – 16	15	6	90	-6,25	234,375	-1,731	-31,136	53,869
16 – 18	17	10	170	-4,25	180,625	-1,177	-16,317	19,191
18 – 20	19	15	285	-2,25	0,9375	-0,623	-3,632	2,259
20 – 22	21	24	504	-0,25	1,5	-0,069	-0,008	0,0005
22 – 24	23	19	437	1,75	58,1875	0,485	2,164	1,0513

Продолжение таблицы 2.3

24 – 26	25	14	350	3,75	196,875	1,038	15,692	16,252
26 – 28	27	6	162	5,75	198,375	1,592	24,245	38,541
28 - 30	29	2	58	7,75	120,125	2,146	19,788	42,417
30 – 32	31	1	31	9,75	95,0625	2,701	19,701	53,222
Итого:	-	100	2125	-	1290,25	-	-5,309	308,58
$\frac{\sum}{100}$			21,25		12,9025		-0,05309	3,0858

Т.о., $\bar{x} = 21{,}25$, $D(X) = 12{,}9025$, $\sigma = \sqrt{D(X)} = \sqrt{12{,}9025} \approx 3{,}59$.

«Исправим» дисперсию: $s^2 = \frac{n}{n-1} \cdot D(X) = \frac{100}{99} \cdot 12{,}9025 = 13{,}04$.

Найдем «исправленное» среднее квадратическое отклонение: $s = \sqrt{13{,}04} = 3{,}61$.

Наряду с абсолютным показателем колеблемости признака - средним квадратическим отклонением, широко применяется и относительный показатель - коэффициент вариации, который показывает меру колеблемости признака относительно его среднего значения и измеряется в процентах. Коэффициент вариации υ определяется по формуле:

$$\upsilon = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{3{,}592}{21{,}25} \cdot 100\% = 17\%$$

Так как коэффициент вариации 10% - 20% то изменчивость признака считается средней. Если коэффициент вариации меньше 10%, то изменчивость считается незначительной, если больше 20% -то значительной.

5. Мода

Для интервального ряда мода находится по формуле:

$$M_0(X) = x_{M_0} + h\frac{(m_2 - m_1)}{(m_2 - m_1) + (m_2 - m_3)}$$

где

x_{M_0} - начало модального интервала;

h - длина частичного интервала;

m_1 - частота предмодального интервала;

m_2 - частота модального интервала;

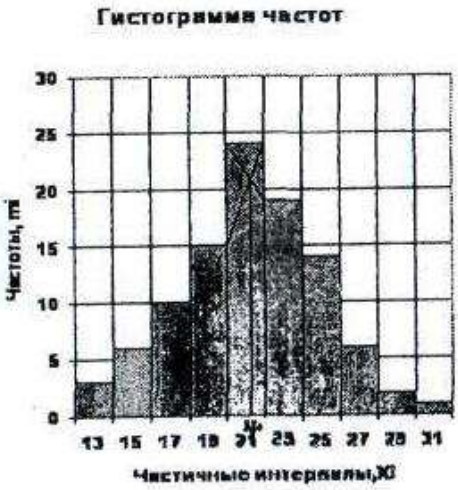
m_3 - частота послемодального интервала.

Определим модальный интервал - интервал, имеющий наибольшую частоту. В *таблице № 2.2* находим, что модальным является интервал 20-22.

Получаем,

$$M_0(X) = 20 + 2 \cdot \frac{(24-15)}{(24-15)+(24-19)} = 21{,}29.$$

Моду можно определить графически по гистограмме вариационного ряда. Для этого правую верхнюю вершину прямоугольника, предшествующего модальному интервалу, соединим прямой линией с противоположной вершиной модального прямоугольника, а левую верхнюю вершину этого прямоугольника- с верхней левой вершиной прямоугольника, следующего за модальным. Из полученной точки пересечения опустить перпендикуляр на горизонтальную ось. Основание этого перпендикуляра и будет модой.



Как видно из графика гистограммы частот, $M_0 = 21{,}3$, что соответствует моде полученной аналитически.

10. Медиана

Для интервального ряда медиана находится по формуле:

$$M_e(X) = x_{M_e} + h\frac{0{,}5n - S_{M_e-1}}{n_{M_e}}$$

где

x_{M_e} - начало медианного интервала;

h - длина частичного интервала;

n - объем совокупности;

S_{M_e-1} - накопленная частота интервала, предшествующая медианному;

n_{M_e} - частота медианного интервала.

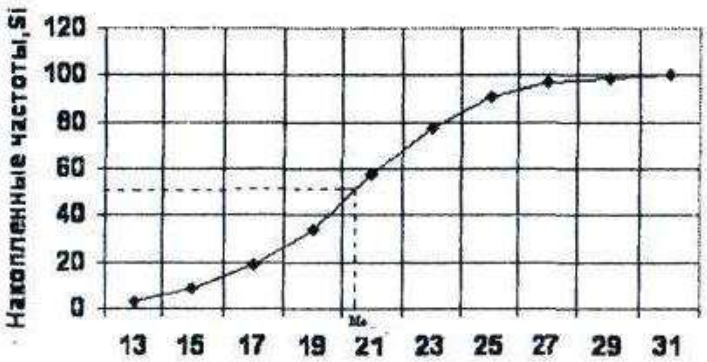
Сначала определим медианный интервал – интервал, в котором *впервые* накопленная частота превышает половину объема выборки. Так как объем выборки n=100, то $n/2$=50. По *таблице № 2.2* найдем интервал, где впервые накопленные частоты превысят это значение. Таким является интервал 20-22.

Получаем,

$$M_e(X) = 20 + 2 \cdot \frac{0{,}5 \cdot 100 - 34}{24} = 21{,}33.$$

Определим теперь медиану графически, используя кумуляту.

Кумулята



Как видно из графика кумуляты, $M_e = 21{,}3$, что соответствует медиане полученной аналитически.

11. Асимметрия и крутизна распределения

Выборочный **коэффициент асимметрии** служит для характеристики асимметрии полигона вариационного ряда. Если полигон асимметричен, то одна из ветвей его начиная с вершины имеет более пологий «спуск», чем другая. Вычислим коэффициент асимметрии Sk по формуле:

$$Sk = \frac{\sum \left(x_i^* - \bar{x}\right)^3 m_i}{n\sigma^3}$$

где k- число интервалов, x_i^* - середины интервалов, $\bar{x}$ - выборочная средняя, m_i- частоты соответствующих интервалов, n - объем выборки, σ - выборочное среднее квадратическое отклонение.

Из предпоследнего столбца *таблицы № 2.3* получаем, $Sk = -0{,}053$. Коэффициент асимметрии меньше нуля - асимметрия отрицательная или левосторонняя, незначительная.

Выборочный эксцесс служит для сравнения на «крутость» выборочного распределения с нормальным. Эксцесс случайной величины, распределенной нормально, равен нулю. Если выборочному распределению соответствует отрицательный эксцесс, то соответствующий полигон имеет более пологую вершину по сравнению с нормальной кривой. В случае положительного эксцесса полигон более крутой по сравнению с нормальной кривой. Коэффициент эксцесса определяется по формуле:

$$Ex = \frac{\sum (x_i - \bar{x})^4 m_i}{n\sigma^4} - 3,$$

где k- число интервалов, x_i^* - середины интервалов, $\bar{x}$ - выборочная средняя, m_i- частоты соответствующих интервалов, n - объем выборки, σ - выборочное среднее квадратическое отклонение.

Из последнего столбца *таблицы № 2.3* получаем, $Ex = 3{,}08 - 3 = 0{,}08$. Выборочный коэффициент эксцесса больше нуля – распределение более крутое, чем нормальное.

2. *Интервальный вариационный ряд*

2.1.Нахождение наблюдаемого значения критерия $\chi^2_{набл}$

Используем данные из лабораторной работы №2: $\bar{x} = 21{,}25 \quad \sigma = 3{,}59$.

В нашем примере имеются интервалы, частота которых меньше 5, поэтому объединим первые два интервала в один и по-

следние три – в один. За начало первого интервала примем $-\infty$, а за конец последнего - $+\infty$.

Теоретические частоты находим по формуле:

$$m_i' = n \cdot p_i,$$

где n - объем выборки, p_i - вероятность попадания в i-й интервал, вычисляется по формуле:

$$p_i = \Phi\left(\frac{x_{i+1} - \bar{x}}{\sigma}\right) - \Phi\left(\frac{x_i - \bar{x}}{\sigma}\right)$$

где $\bar{x}$ - выборочная средняя, σ - выборочное среднее квадратическое отклонение, x_i и x_{i+1} - соответственно начало и конец i-го интервала, $\Phi(x)$ - интегральная функция Лапласа.

Замечания.

1. Функция $\Phi(u)$ является интегральной функцией Лапласа, значения которой можно найти по таблицам приложения 2.
3. Функция $\Phi(u)$ является нечетной, т.е.. $\Phi(-x) = -\Phi(x)$

Результаты оформим в виде таблицы.

Таблица 3.3.

Интервалы $[x_i; x_{i+1})$	Эмпирические частоты m_i	$p_i = \Phi(u_{i+1}) - \Phi(u_i)$	Теоретические частоты, m_i'	$(m_i - m_i')^2$	$\frac{(m_i - m_i')^2}{m_i'}$
$-\infty$ - 16	9	0,0721	7,21	3,2041	0,4443
16 – 18	10	0,1093	10,93	0,8649	0,0791
18 – 20	15	0,1855	18,55	12,6025	0,6793
20 – 22	24	0,2163	21,63	5,6169	0,2597
22 – 24	19	0,1962	19,62	0,3844	0,0196
24 – 26	14	0,1272	12,72	1,6384	0,1288
26 - $+\infty$	9	0,0934	9,34	0,1156	0,0124
$\sum$	100	1	100		$\chi^2_{расч} = 1,62$

$$p_1 = F(-\infty < x \le 16) = \Phi\left(\frac{16 - 21,25}{3,59}\right) - \Phi\left(\frac{-\infty - 21,25}{3,59}\right) = \Phi(-1,46) - \Phi(-\infty) = -\Phi(1,46) + \Phi(\infty) =$$

$$= -0,4279 + 0,5 = 0,0721;$$

$$p_2 = F(16 < x \le 18) = \Phi\left(\frac{18 - 21,25}{3,59}\right) - \Phi\left(\frac{16 - 21,25}{3,59}\right) = \Phi(-0,91) - \Phi(-1,46) = -\Phi(0,91) + \Phi(1,46) =$$

$$= -0,3186 + 0,4279 = 0,1093;$$

и т.д.

Итак, мы нашли наблюдаемое значение критерия $\chi^2_{набл} = 1,62$.

2.2.Нахождение критических точек распределения χ^2

Найдем число степеней свободы k но формуле:

$$k = s - r - 1,$$

где s - число интервалов после объединения, r - число параметров, оцениваемых по выборке.

Итак

$$k = 7 - 2 - 1 = 4.$$

По таблице критических точек распределения χ^2 (приложение 3), по заданному уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $k = 4$ находим критическую точку $\chi^2_{табл}(\alpha, k)$:

$$\chi^2_{табл}(0{,}05;4) = 9{,}5.$$

Так как $\chi^2_{набл} < \chi^2_{табл}$ - гипотезу о нормальном распределении генеральной совокупности принимаем. Расхождение между эмпирическими и теоретическими частотами незначимо (случайно).