ГЛАВА 2 ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

2.1 Предмет и задачи математической статистики.Вариационные ряды

Математическая статистика используется в различных областях знаний: в экономике, опытном деле, земледелии, животноводстве и т.д., т.е. там, где для изучения процессов и явлений недостаточно только качественной характеристики. Чтобы глубоко познать сущность процессов, необходимы количественные характеристики в виде измерений, наблюдений с их последующим анализом, обобщением и выводами.

Математическая статистика–это наука, занимающаяся разработкой методов сбора, регистрации и обработки результатов наблюдений (измерений) с целью познания закономерностей случайных массовых явлений.

Результаты измерений (наблюдений) называют ***статистическими данными***.

Вся исследуемая совокупность однородных объектов называется ***генеральной совокупностью***.

Множество из ***n***-объектов, отобранных случайным образом из генеральной совокупности, называется ***выборочной совокупностью*** или ***выборкой*** (***n***-объем выборки).

Одним из основных способов сбора статистических данных является ***выборочный метод.***

Метод, основанный на том, что по данным обследования выборки, выделенной из данной генеральной совокупности, делается заключение обо всей генеральной совокупности, называется ***выборочным методом***.

Значение случайной величины, соответствующее отдельной группе сгруппированного ряда наблюдаемых данных, называется ***вариантом*** (x_i), а изменения этого значения – ***варьированием.***

Результаты наблюдений, в общем случае– рядчисел, расположены в беспорядке, поэтому их необходимо упорядочить.

Вариационным рядомназывается ранжирование в порядке возрастания вариант с соответствующими им частотами (ранжир – впереводе с фр.- «ставить в ряд по росту»).

Численность отдельной группы сгруппированного ряда наблюдаемых данных называется ***частотой*** или ***весом*** соответствующего варианта и обозначается m_i, где i - индекс варианта.

Отношение частоты данного варианта к объему совокупности называется ***относительной частотой*** (w_i) или***частостью*** этого варианта.

$$\boxed{w_i = \frac{m_i}{n}}$$

Дискретным вариационным рядом распределения называется ранжированная совокупность вариантов x_i с соответствующими им частотами m_i или частостями (w_i).

В общем виде его можно записать так:

x_i	x_1	x_2	…	x_n
m_i	m_1	m_2	…	m_n

Вариационный ряд часто дополнительно характеризуется накопленными частотами или накопленными частостями.

Накопленные частоты характеризуют число членов данной совокупности, у которых рассматриваемый признак принимает значения, не превышающие данного варианта.

Накопленные частости – результаты последовательного суммирования частостей всех вариантов, включая частость данного варианта.

Кроме дискретных вариационных рядов широкое применение имеют ***непрерывные (интервальные)*** вариационные ряды.

Интервальным вариационным рядом называется упорядоченная совокупность интервалов варьирования значений случайной величины с соответствующими частотами или частостями попаданийв каждый из них значений случайной величины.

Интервальный ряд целесообразно построить, если число возможных значений дискретной величины велико или признак является непрерывным, т.е. может принимать любые значения в пределах некоторого интервала.

Для построения интервального ряда необходимо определить величину частичных интервалов, на которые разбивается весь интервал варьирования наблюдаемых значений случайной величины. Считая, что все частичные интервалы имеют одну и ту же длину, для каждого интервала следует установить его верхнюю и нижнюю границы, а затем в соответствии с полученной упорядоченной совокупностью частичных интервалов сгруппировать результаты наблюдений.

Для определения величины частичного интервала воспользуемся формулой Стерджесса:

$$\boxed{h = \frac{X_{\max} - X_{\min}}{1 + 3{,}322 \lg n}}$$, где $\boxed{1 + 3{,}322 \lg n}$ -число интервалов.

За начало первого интервала рекомендуется брать величину:

$$\boxed{X_{нач.} = X_{\min} - \frac{h}{2}}$$

Промежуточные интервалы получают, прибавляя к концу предыдущего интервала длину частичного интервала ***h***.

Теперь, просматривая, результаты наблюдений, определяем, сколько значений признака попало в каждый конкретный интервал. При этом в интервал включают значение случайной величины, большие или равные нижней границе и меньшие верхней границы.

Пример.Пусть дан ряд распределения хозяйств по количеству рабочих на 100 га с/х угодий $(n=60)$:

12	6	8	6	10	11	7	10	12	8	7	7	6	7	8	6	11	9	11	9	10
11	9	10	7	8	8	8	11	9	8	7	5	9	7	7	14	11	9	8	7	4
7	5	5	10	7	7	5	8	10	10	15	10	10	13	12	11	15	6			

Построить интервальный вариационный ряд.

Решение.Для определения числа групп подставим значение $n=60$ в формулу Стерджесса:

$$k=1+3,322\lg 60\approx 6,907;\quad k=7$$

Найдем длину частичного интервала: $h=\dfrac{X_{\max}-X_{\min}}{k}=\dfrac{15-4}{7}=\dfrac{11}{7}\approx 1,6$

Построим интервальный вариационный ряд, для этого в качестве начального значения используем $X_{\min}$.

Группы хозяйств по численности работников на 100 га с/х угодий	Число хоз-в в группе, m_i	Накопленное число хоз-в, S_i	Относительная частота, $\widehat{p}_i$
4- 5,6	5	5	5/60
5,61- 7,2	17	22	17/60
7,21- 8,8	9	31	9/60
8,81- 10,4	15	46	15/60
10,41- 12,0	10	56	10/60
12,01- 13,6	1	57	1/60
13,61- 15,2	3	60	3/60
Итого	60	-	1

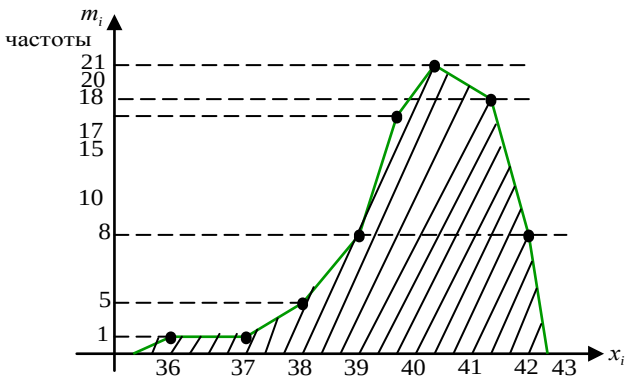
Иногда интервальный вариационный ряд для простоты исследований условно заменяют дискретным.

В этом случае серединное значение $i-го$ интервала принимают за вариант x_i, а соответствующую интервальную частоту m_i- за частоту этого интервала.

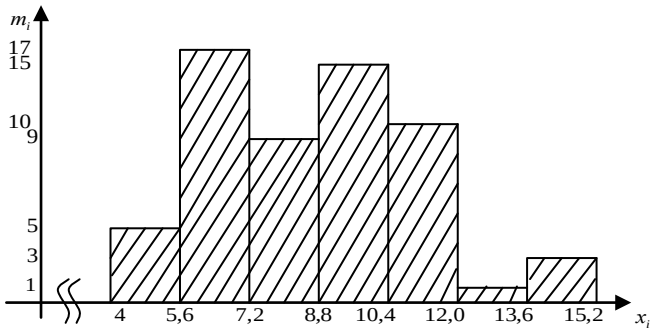
2.2 Графическое изображение вариационных рядов

Графическое изображение позволяет представить в наглядной форме закономерности варьирования значений признаков с помощью полигона, гистограммы, кумуляты и огивы.

Полигоном (для дискретного вариационного ряда) называется ломанная, соединяющая на плоскости точки с координатами $(x_i;m_i)$.



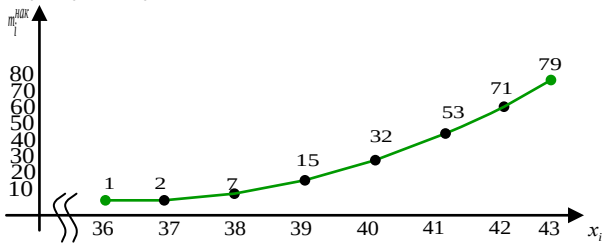
Гистограммой (для интервального вариационного ряда) называют ступенчатую фигуру, состоящую и прямоугольников, основаниями которых служат интервалы $(x_{i-1};x_i)$, а высотами – частоты m_i.



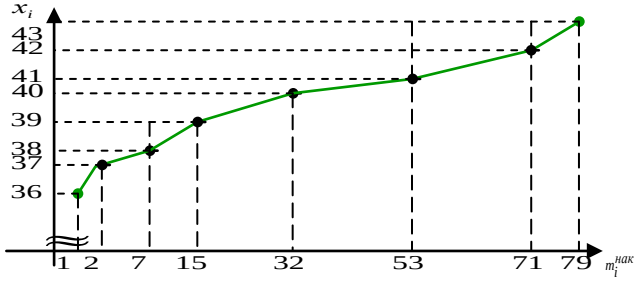
Если в вариационном ряду вместо частот взяты соответственно накопленные частоты, то полученный ряд называется ***кумулятивным рядом*** (кумуляция – от латинского «скопление»).

Кумулятойназывается ломанная, соединяющая на плоскости точки вида $(x_i;m_i^{нак})$.

Кумулятуиначе называют полигоном накопленных частот.



Если по оси абсцисс откладывать накопленные частоты, а по оси ординат - значение признака, затем полученные точки соединить отрезками, то получится ***огива.***



2.3 Числовые характеристики вариационных рядов

Вариационные ряды позволяют получить первое представление об изучаемом распределении. Далее необходимо исследовать числовые характеристики распределения (аналогичные характеристикам распределения теории вероятностей): характеристики **положения** (средняя арифметическая, мода, медиана); характеристики **рассеивания** (дисперсия, среднее квадратическое отклонение, коэффициент вариации); характеристики **меры скошенности** (коэффициент асимметрии) и **островершинности** (эксцесс) распределения.

Средней арифметической $\left(\overline{X}\right)$ дискретного вариационного ряда называется отношение суммы произведений вариантов на соответствующие частоты к объему совокупности:

$$\overline{X} = \frac{\sum x_i m_i}{n} \qquad (38)$$

Вычисленное по формуле (38)среднее арифметическое называется **взвешенным**, так как частоты m_i называются**весами,** а операция умножения x_i на m_i - **взвешиванием.**

Для интервального вариационного ряда за x_i принимают середину $i-го$ интервала, а за m_i - соответствующую интервальную частоту.

Модой $\left(\widehat{M}_0\left(x\right)\right)$ дискретного вариационного ряда называется вариант, имеющий наибольшую частоту.

Для интервальных вариационных рядов при нахождении $\widehat{M}_0\left(x\right)$ используют формулу:

$$\boxed{\widehat{M}_0\left(x\right) = x_0 + h \cdot \frac{m_i - m_{i-1}}{\left(m_i - m_{i-1}\right) + \left(m_i - m_{i+1}\right)}},$$

где x_0 - начало модального интервала;

h - длина частичного интервала;

m_i - частота модального интервала;

m_{i-1} - частота предмодального интервала;

m_{i+1} - частота послемодального интервала.

Медианой $\left(\widehat{M}_e\left(x\right)\right)$ дискретного вариационного ряда называется вариант, делящий ряд на две равные части.

Если дискретный вариационный ряд имеет **четное** $\left(2n\right)$ число членов, то:

$$\boxed{\widehat{M}_e\left(x\right) = \frac{x_n + x_{n+1}}{2}}.$$

Если дискретный вариационный ряд имеет **нечетное** $\left(2n-1\right)$ число значений варьирующего признака, расположенных в порядке возрастания, то медианой этого распределения является вариант x_n:

$$\widehat{M}_e(x) = x_n$$

При нахождении $\widehat{M}_e(x)$ для интервальных вариационных рядов используют формулу:

$$\widehat{M}_e(x) = x_0 + h \cdot \frac{0{,}5n - m_{i-1}^{нак}}{m_i},$$

где x_0 - начало медианного интервала;

h - длина частичного интервала; n - объем совокупности;

$m_{i-1}^{нак}$ - накопленная частота интервала, предшествующего медианному;

m_i - частота медианного интервала.

Пример. Найти $\widehat{M}_e(x)$ по условию задачи пункта 3.1.

$n = 60 \Rightarrow \frac{n}{2} = 30 \Rightarrow$ медиана расположена в интервале (7,21;8,8).

$$\widehat{M}_e(x) = 7{,}21 + 1{,}6\frac{0{,}5 \cdot 60 - 22}{9} \approx 8{,}62$$

Дисперсия вариационного ряда (как дискретного, так и интервального) характеризует средний квадрат отклонения значения признака от его среднего значения: $\widehat{D}(x) = \frac{\sum (x_i - \bar{x})^2 \cdot m_i}{n}$

Среднее квадратическое отклонение вариационного ряда распределения характеризует те же значения, что и дисперсия, но измеряется в единицах варьирующего признака: $\hat{\sigma}(x) = \sqrt{\frac{\sum (x_i - \bar{x})^2 \cdot m_i}{n}}$

Коэффициент вариации характеризует относительное значение среднего квадратического отклонения и служит для сравнения колеблемости несоизмеримых показателей: $V = \frac{\hat{\sigma}(x)}{\bar{X}} \cdot 100\%$

Коэффициент асимметрии - $\widehat{A} = \frac{\sum (x_i - \bar{x})^3 \cdot m_i}{n \cdot \hat{\sigma}^3(x)}$

Эксцесс - $\widehat{E} = \frac{\sum (x_i - \bar{x})^4 \cdot m_i}{n \cdot \hat{\sigma}^4(x)} - 3$

Пример: Рассчитать дисперсию, среднее квадратическое отклонение, коэффициенты вариации, асимметрии и эксцесс для задачипункта 3.1. Сделать выводы.

Решение.Построим вспомогательную таблицу.

Группы хоз-в по численности	Среднее значение интерва	Число хоз-в в группе,	$x_i \cdot m_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 \cdot m_i$	$\frac{x_i - \bar{x}}{\hat{\sigma}(x)}$	$\left(\frac{x_i - \bar{x}}{\hat{\sigma}(x)}\right)^3 \cdot m_i$	$\left(\frac{x_i - \bar{x}}{\hat{\sigma}(x)}\right)^4 \cdot m_i$

работников на 100 га с/х угодий, чех.	ла, x_i	m_i						
4-5,6	4,8	5	24	-3,813	72,708	-1,559	-18,954	29,554
5,61-7,2	6,4	17	108,8	-2,213	83,280	-0,905	-12,601	11,404
7,21-8,8	8	9	72	-0,613	3,386	-0,251	-0,142	0,036
8,81-10,4	9,6	15	144	0,987	14,603	0,403	0,985	0,397
10,41-12,0	11,2	10	112	2,587	66,908	1,058	11,832	12,514
12,01-13,6	12,8	1	12,8	4,187	17,528	1,712	5,017	8,588
13,61-15,2	14,4	3	43,2	5,787	100,457	2,366	39,740	94,030
Итого	-	60	516,8	0	358,869	-	25,876	156,523

$$\bar{x} = \frac{516,8}{60} = 8,613$$

$$\hat{D}(x) = \frac{\sum\left(x_i - \bar{x}\right)^2 \cdot m_i}{n} = \frac{358,869}{60} = 5,981$$

$$\hat{\sigma}(x) = \sqrt{\hat{D}(x)} = \sqrt{5,981} \approx 2,446$$

$$V = \frac{\hat{\sigma}(x)}{\bar{x}} \cdot 100\% = \frac{2,446}{8,613} \cdot 100\% = 28,4\%$$

Таким образом, средняя численность работников на 100 га с/х угодий по исследуемой совокупности хозяйств составила 8,61 чел. Плотность работников в среднем колебалась в промежутке $\boxed{\bar{x} \pm \hat{\sigma}(x)}$ = 8,61±2,45, т.е. от 6,16 до 11,06 чел. на 100 га с/х угодий.

Этот интервал, а так же коэффициент вариации показывает, что имеются большие различия в обеспечении хозяйств рабочей силой.

$$\hat{A} = \frac{\sum\left(x_i - \bar{x}\right)^3 \cdot m_i}{n \cdot \hat{\sigma}^3(x)} = \frac{25,876}{60} = 0,43$$

$$\hat{E} = \frac{\sum\left(x_i - \bar{x}\right)^4 \cdot m_i}{n \cdot \hat{\sigma}^4(x)} - 3 = \frac{156,523}{60} - 3 = -0,39$$

Найденное значение коэффициента асимметрии (не достаточно близкое к нулю) указывает, что распределение не симметрично. Эксцесс также отличен от нуля, что говорит о возможном отличии распределения от нормального.

2.4 Выборочный метод. Точечные и интервальные оценки параметров распределения

В реальных условиях обычно бывает трудно или экономически нецелесообразно, а иногда и невозможно, исследовать всю совокупность, характеризующую изучаемый признак (генеральную совокупность). Поэтому на практике широко применяется выборочное наблюдение, когда обрабатывается часть генеральной совокупности (выборочная совокупность). Свойства (закон распределения и его параметры) генеральной совокупности неизвестны, поэтому возникает задача их оценки по выборке. Для получения

хороших оценок характеристик генеральной совокупности необходимо, чтобы выборка была репрезентативной (представительной). Репрезентативность в силу закона больших чисел, достигается случайностью отбора.

После осуществления выборки возникает задача оценки числовых характеристик генеральной совокупности по элементам выборочной совокупности. Различают точечные и интервальные оценки.

Точечной оценкой характеристики генеральной совокупности называется число, определяемое по выборке.

Пусть $\widehat{\Theta} = \widehat{\Theta}_n$ выборочная характеристика, вычисленная по результатам n наблюдений величины X, используемая в качестве оценки Θ-характеристики генеральной совокупности (в качестве Θ может быть $M(x)$; $D(x)$ и т.д.).

Качество оценки $\widehat{\Theta}$ устанавливается по 3-м свойствам:

1. **Состоятельность.**Оценка $\widehat{\Theta}_n$ является состоятельной оценкой генеральной совокупности Θ, если для любого $\varepsilon > 0$ выполняется неравенство:

$$\boxed{\lim_{n\to\infty} P\left(\left|\widehat{\Theta}_n - \Theta\right| < \varepsilon\right) = 1}$$

Это означает, что при увеличении объема выборки **n** выборочная характеристика стремится к соответствующей характеристике генеральной совокупности $\left(\widehat{\Theta}_n \to \Theta\right)$.

2. **Несмещенность.**Оценка $\widehat{\Theta}_n$ генеральной характеристики Θ называется несмещенной, если для любого фиксированного числа наблюдений **n** выполняется равенство:

$$\boxed{M\left(\widehat{\Theta}_n\right) = \Theta}$$

3. **Эффективность.** Несмещенная оценка $\widehat{\Theta}_n$ генеральной характеристики Θ называется несмещенной эффективной, если среди всех подобных оценок той же характеристики она имеет наименьшую дисперсию:

$$\boxed{D\left(\widehat{\Theta}_n\right) \to \min}$$

Статистики $\overline{x}$ и $\widehat{p}_i$ являются состоятельными, несмещенными и эффективными характеристиками математического ожидания $M(x)$ и вероятности P соответственно.

Выборочная дисперсия $D(x) = \hat{\sigma}^2(x)$ не обладает свойством несмещенности. На практике используют **исправленную выборочную дисперсию** S^2**,** которая является несмещенной оценкой дисперсии генеральной совокупности:

$$S^2 = \frac{n}{n-1}\cdot\sigma^2(x) = \frac{n}{n-1}\cdot\frac{\sum\left(x_i-\bar{x}\right)^2\cdot m_i}{n} = \frac{\sum\left(x_i-\bar{x}\right)^2\cdot m_i}{n-1} \Rightarrow$$

$$\boxed{S^2 = \frac{\sum\left(x_i-\bar{x}\right)^2\cdot m_i}{n-1}},$$

где S – стандартное отклонение.

Интервальной называют оценку, которая определяется двумя числами- границами интервала.

Интервальная оценка позволяет ответить на вопрос: внутри какого интервала, и с какой вероятностью находится неизвестное значение оцениваемого параметра генеральной совокупности?

Пусть $\hat{\Theta}$ точечная оценка параметра Θ. Чем меньше разность $\hat{\Theta}-\Theta$, тем точнее и лучше оценка. Обычно говорят о **доверительной вероятности** $\boxed{p=1-\alpha}$, с которой Θ будет находиться в интервале:

$$\boxed{\hat{\Theta}-\Delta<\Theta<\hat{\Theta}+\Delta},$$

где $\Delta\left(\Delta>0\right)$- предельная ошибка выборки, которая может быть задана наперед, либо вычислена; α - риск или уровень значимости (вероятность того, что неравенство будет неверным).

В качестве $\left(1-\alpha\right)$ принимают значения 0,90; 0,95; 0,99; 0,999. Доверительная вероятность показывает, что в $\left(1-\alpha\right)\cdot 100\%$ случаев оценка будет накрываться указанным интервалом.

Точечная оценка математического ожидания $M\left(x\right)=a$ определяется как средняя арифметическая:

$$\boxed{\bar{x}=\frac{1}{n}\sum x_i\cdot m_i}.$$

Для построения доверительного интервала параметра a- математического ожидания нормального распределения, составляют выборочную характеристику (статистику), функционально зависимую от наблюдений и связанную с a, например:

1. для **повторного отбора:**

$$\boxed{u=\frac{\bar{x}-a}{\frac{\sigma(x)}{\sqrt{n}}}}$$

Статистика **u** распределена по нормальному закону распределения с математическим ожиданием $a=0$ и средним квадратическим отклонением $\sigma(x)=1$. Отсюда:

$$\boxed{P\left(|u|<U_{\alpha/2}\right)=1-\sigma(x)} \text{ или } \boxed{2\Phi\left(U_{\alpha/2}\right)=1-\sigma(x)},$$

где Ф- функция Лапласа.

$U_{\alpha/2}$ - квантиль нормального закона распределения, соответствующая уровню значимости α.

Доверительный интервал для параметра a:

$$\boxed{\bar{x} - U_{\alpha/2} \cdot \frac{\sigma(x)}{\sqrt{n}} < a < \bar{x} + U_{\alpha/2} \cdot \frac{\sigma(x)}{\sqrt{n}}}$$

2. для **бесповторного отбора**:

Доверительный интервал для средней:

$$\boxed{\bar{x} - \Delta_{\bar{x}} < \overline{x_0} < \bar{x} + \Delta_{\bar{x}}},$$

где $\bar{x}$ - выборочная средняя;

$\overline{x_0}$ - средняя генеральной совокупности;

$\Delta_{\bar{x}}$ - предельная ошибка выборки для средней.

Предельная ошибка выборки:

$$\boxed{\Delta_{\bar{x}} = t \cdot \sqrt{\frac{S^2}{n}\left(1 - \frac{n}{N}\right)}},$$

где t - квантиль нормального закона распределения (при $\alpha = 0,05 \quad t = 1,96$);

N- объем генеральной совокупности;

n- объем выборки;

S^2 - исправленная выборочная дисперсия.

Квантилем или ***нормированным отклонением*** называется отношение предельной ошибки к средней ошибке.

$$\boxed{t = \frac{\Delta_{\bar{x}}}{M_{\bar{x}_0}}},$$

где $M_{\bar{x}_0} = \frac{\sigma(x)}{\sqrt{n}}$

Квантиль вычисляется по соответствующему уровню значимости α (при $n \geq 30$, t - квантиль нормального закона распределения, при $n < 30$, t - квантиль распределения Стьюдента).

Существуют таблицы значений для t в зависимости от уровня значимости α.

Важной является задача определения объема выборочной совокупности ***n*** при заданном уровне значимости. В случае бесповторного отбора необходимый объем выборки определяется по формуле:

$$\boxed{n = \frac{t^2 \cdot S^2 \cdot N}{t^2 \cdot S^2 + \Delta_{\bar{x}}^2 \cdot N}}$$

Пример.По условию задачи пункта 3.1. При уровне значимости $\alpha = 0,05$ определить:

1) несмещенные оценки математического ожидания, дисперсии и среднего квадратического отклонения;

2) доверительный интервал для математического ожидания с доверительной вероятностью $(1 - \alpha)$;

3) объем выборки, при котором с доверительной вероятностью $(1-\alpha)$ предельная ошибка выборки уменьшится в 2 раза, при сохранении уровня остальных характеристик.

Учитывая, что проводилась 10% случайная бесповторная выборка.

Решение.

1) Несмещенной оценкой $M(x)$ является выборочная средняя $\bar{x}$:

$$\bar{x} = 8,613$$

Несмещенной оценкой $D(x)$ является исправленная выборочная дисперсия S^2:

$$S^2 = \sigma^2(x) \cdot \frac{n}{n-1} = \frac{5,981 \cdot 60}{59} = 6,082$$

Несмещенной оценкой $\sigma(x)$ является стандартное отклонение S:

$$S = \sqrt{S^2} = \sqrt{6,082} = 2,466$$

2) Средняя численность работников на 100 га с/х угодий = 8,61. Найдем доверительный интервал для средней: $\bar{x} - \Delta_{\bar{x}} < \overline{x_0} < \bar{x} + \Delta_{\bar{x}}$

$\Delta_{\bar{x}} = t \cdot \sqrt{\frac{S^2}{n}\left(1 - \frac{n}{N}\right)}$ - предельная ошибка выборки для средней.

При уровне значимости $\alpha = 0,05$ квантиль нормального распределения $t = 1,96$.

Учитывая, что проводилась 10% выборка,

$$N = 10 \cdot 60 = 600 \quad \Rightarrow \Delta_{\bar{x}} = 1,96 \cdot \sqrt{\frac{6,082}{60}\left(1 - \frac{60}{600}\right)} = 0,592$$

Значит, с доверительной вероятностью $1-\alpha = 0,95$, можно утверждать, что средняя численность работников на 100 га с/х угодий во всей совокупности хозяйств находится в границах $\bar{x} \pm \Delta_{\bar{x}} = 8,61 \pm 0,592$, т.е. от 8,021 до 9,205.

3) Необходимый объем выборки, для того, чтобы предельная ошибка не превышала $0,5 \cdot \Delta_{\bar{x}}$ при заданном уровне значимости $\alpha = 0,05$ в случае случайного бесповторного отбора, определяется по формуле:

$$n = \frac{t^2 \cdot S^2 \cdot N}{t^2 \cdot S^2 + \Delta_{\bar{x}}^2 \cdot N},$$

$$\Delta_{\bar{x}}^2 = \left(0,5 \cdot \Delta_{\bar{x}}\right)^2 = \left(0,5 \cdot 0,592\right)^2 = \left(0,296\right)^2$$

$$n = \frac{(1,96)^2 \cdot 6,082 \cdot 600}{(1,96)^2 \cdot 6,082 + (0,296)^2 \cdot 600} = \frac{14018,766}{23,365 + 52,576} = 185$$

Значит, для уменьшения предельной ошибки в два раза объем совокупности необходимо увеличить в 3 раза.

Задания длясамостоятельного решения

1.При изменении диаметра валика после шлифовки была получена следующая выборка (объемом $n = 55$):

20.3	15.4	17.2	19.2	23.3	18.1	21.9
15.3	16.8	13.2	20.4	16.5	19.7	20.5
14.3	20.1	16.8	14.7	20.8	19.5	15.3
19.3	17.8	16.2	15.7	22.8	21.9	12.5
10.1	21.1	18.3	14.7	14.5	18.1	18.4
13.9	19.8	18.5	20.2	23.8	16.7	20.4
19.5	17.2	19.6	17.8	21.3	17.5	19.4
17.8	13.5	17.8	11.8	18.6	19.1	

Необходимо: 1)составить интервальный вариационный ряд, построить полигон и гистограмму; 2) найти моду и медиану; 3) рассчитать дисперсию, среднее квадратическое отклонение, коэффициенты вариации, асимметрии и эксцесс. Сделать выводы.

При уровне значимости $\alpha = 0,05$ определить:

1) несмещенные оценки математического ожидания, дисперсии и среднего квадратического отклонения;

2) доверительный интервал для математического ожидания с доверительной вероятностью $(1-\alpha)$;

3) объем выборки, при котором с доверительной вероятностью $(1-\alpha)$ предельная ошибка выборки уменьшится в 2 раза, при сохранении уровня остальных характеристик.

Учитывая, что проводилась 10% случайная бесповторная выборка.

2.5 Элементы теории корреляции

В сельскохозяйственных науках, в отличие от точных наук, полные (точные) функциональные связи встречаются редко, так как возможность искусственной изоляции влияния других факторов на изучаемые признаки в большинстве случаев неосуществима.

Например, связь урожайность– удобрения, имеется,но есть еще ряд факторов, влияющих на урожайность(севообороты, семена, предшественники, агротехника– субъективные факторы; метеорологические факторы- объективные).

Поэтому связь урожайность– удобрения неполная функциональная связь. Эту связь называют **корреляционной**(англ. correlation – соотношение, соответствие).

Метод корреляции применяется для того, чтобы при сложном взаимодействии посторонних влияний выяснить, какова была бы зависимость между результатом и фактором, если бы посторонние причины

(факторы) не изменялись и своим изменением не искажали основную зависимость.

Первая задача корреляции: выявление на основе наблюдений над большим количеством фактов того, как изменяется в среднем результативный признак в связи с изменением данного фактора (парная корреляция) или группы факторов (множественная корреляция). Эта задача решается нахождением уравнения связи.

Вторая задача корреляции: определение степени влияния искажающих факторов. Эта задача решается при помощи различных показателей тесноты связи: коэффициента корреляции, корреляционного отношения.

Процесс нахождения связи между признаками называется ***выравниванием.***

Выравнивание ведет к нахождению переменной средней $\overline{y}_x$, исчисленной в предположении функциональной зависимости **y** от **x**, т.е. $\overline{y}_x = f(x)$, и называется **уравнением регрессии.**

При изучении влияния одних признаков на другие выделяются два признака –**факториальный** и **результативный**. Выделение этих признаков осуществляется путем логического анализа.

Например,в связи урожайность – осадки, урожайность – результативный признак, а осадки – факториальный.

Графическое изображение связи

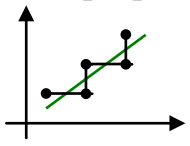
Графическое изображение связи изучаемых явлений позволяет не только установить наличие или отсутствие связи между ними, но и изучить характер этой связи (форму связи и тесноту связи).

Если имеются числовые характеристики факториальных и результативных признаков одного и того же явления, то каждую пару чисел можно изобразить графически, откладывая по оси абсцисс– факториальный признак, по оси ординат– результативный признак.

Ломаная, соединяющая эти точки, называется **ломаной регрессии.**

По форме этой ломаной приближенно определяют вид зависимости.

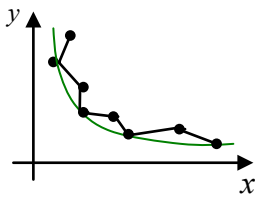
1. Если из графика видно, что связь близка к прямолинейной, то уравнение регрессии пишется в виде:



$$\boxed{\overline{y} = ax + b}$$

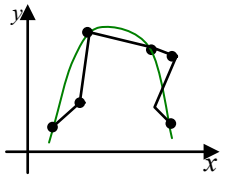
2. Если экспериментальные данные располагаются так, что через них можно провести гиперболу, то можно ожидать уравнение в виде:

$$\boxed{\overline{y} = \frac{k}{x};\quad \overline{y} = \frac{a}{x+b},\quad \overline{y} = \frac{a}{x+b} + c}$$



3. Если кривая имеет max или min, то зависимость определяется по уравнению:

$$\boxed{\overline{y} = ax^2 + bx + c}$$



Для выявления функциональных зависимостей и определения неизвестных коэффициентов этой зависимости можно воспользоваться методом наименьших квадратов:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i \end{cases} \Rightarrow \quad y = ax + b$$

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 \cdot y \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \cdot y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + c \cdot n = \sum_{i=1}^{n} y_i \end{cases} \Rightarrow \quad y = ax^2 + bx + c$$

<u>Коэффициент корреляции</u>

После того, как уравнение регрессии найдено, находят так называемый ***коэффициент корреляции.*** Он используется для оценки тесноты связи между величинами при прямолинейной зависимости. Обозначается буквой r и определяется по формуле:

$$\boxed{r = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\cdot\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}},$$

где $\overline{x}$ – среднее значение факториального (причинного) признака $\overline{x} = \frac{\sum x_i}{n}$

$\overline{y}$ – среднее значение результативного признака $\overline{y} = \frac{\sum y_i}{n}$

Промежуточные вычисления удобно располагать в виде таблицы:

№наблюден	x_i	y_i	$x_i - \overline{x}$	$\left(x_i - \overline{x}\right)^2$	$y_i - \overline{y}$	$\left(y_i - \overline{y}\right)^2$	$\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$

ия							
$\sum$	…	…	…	…	…	…	…

Величина коэффициента корреляции находится в пределах $\boxed{-1 \le r \le 1}$:

1) чем ближе $|r|$ к 1, тем теснее связь между факториальным и результативным признаками.

2) при $|r| = 1$ получается полная функциональная связь.

3) если $|r| \to 0$, то связь между признаками слабая.

4) при $|r| = 0$ связи между признаками нет (линейная зависимость отсутствует).

5) при $r > 0$ зависимость между признаками прямая (возрастающая).

6) при $r < 0$ зависимость обратная (убывающая).

Если зависимость между признаками прямая, то можно пользоваться уравнением прямой регрессии: $\boxed{y - \overline{y} = b_{y/x}\left(x - \overline{x}\right)}$, где $b_{y/x}$ – коэффициент регрессии, который определяется по формуле:

$$\boxed{b_{y/x} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\cdot\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}$$

Пример. Для 10 петушков леггорнов 15 дневного возраста были получены следующие данные о весе их тела (***x***) в граммах и весе гребня (**y**) (в мг):

x_i	83	72	69	90	90	95	95	91	75	70
y_i	56	42	18	84	56	107	90	68	31	48

Требуется:

1) найти коэффициент корреляции и сделать вывод о тесноте и направлении линейной корреляционной связи между признаками;

2) составить уравнение прямой регрессии;

3) нанести на чертеж исходные данные и построить прямую регрессии.

Решение. Составим вспомогательную таблицу:

№	x_i	y_i	$\left(x_i - \overline{x}\right)$	$\left(x_i - \overline{x}\right)^2$	$y_i - \overline{y}$	$\left(y_i - \overline{y}\right)^2$	$\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$
1	83	56	0	0	-4	16	0
2	72	42	-11	121	-18	324	198
3	69	18	-14	186	-42	1764	588
4	90	84	7	49	24	576	168
5	90	56	7	49	-4	16	-28
6	95	107	12	144	47	2209	564
7	95	90	12	144	30	900	360
8	91	68	8	64	8	64	64

9	75	31	-8	64	-29	841	232
10	70	48	-13	169	12	144	156
$\sum$	30	600	0	990	0	6854	2302

Вычисляем средние:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{830}{10} = 83 \quad \bar{y} = \frac{\sum y_i}{n} = \frac{600}{10} = 60$$

1) найдем коэффициент корреляции:

$$r = \frac{\sum \left(x_i - \bar{x}\right) \cdot \left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2} \cdot \sqrt{\sum \left(y_i - \bar{y}\right)^2}} \quad r = \frac{2302}{\sqrt{990 \cdot 6854}} = 0{,}88$$

*Вывод:*между весом тела ***x*** и весом гребня ***y***у 15-тидневных петушков существует тесная положительная линейная корреляционная связь.

2) найдем коэффициент регрессии:

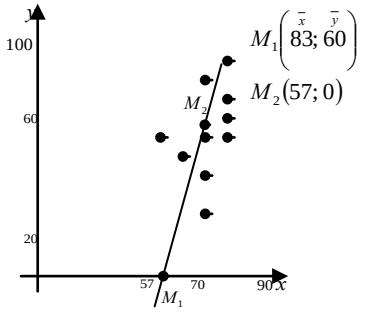
$$b_{y/x} = \frac{\sum \left(x_i - \bar{x}\right) \cdot \left(y_i - \bar{y}\right)}{\sum \left(x_i - \bar{x}\right)^2} \quad b_{y/x} = \frac{2302}{990} \approx 2{,}32$$

Подставим в уравнение прямой регрессии:

$$y - \bar{y} = b_{y/x}\left(x - \bar{x}\right) \quad y - 60 = 2{,}32\left(x - 83\right)$$

$$\underline{y = 2{,}32x - 132{,}56}$$

3) наносим исходные данные на координатную плоскость и строим найденную прямую регрессии.



Задания длясамостоятельного решения

Дана таблица значений *хиу*

x	2,8	3,4	3,7	3,4	2,8	1,5	4,9	7,2	1,7	3,4
y	1,3	2,0	4,4	3,0	2,2	1,8	5,0	2,8	9,1	4,4

Требуется:

1. найти коэффициент корреляции и сделать вывод о тесноте и направлении линейной корреляционной связи между признаками;
2. составить уравнение прямой регрессии;
3. нанести на чертеж исходные данные и построить прямую регрессии.