ЧУДНОВСКАЯ Г.В.

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В БИОЛОГИИ

УЧЕБНОЕ ПОСОБИЕ

Иркутск 2012

Печатается по решению методического совета Иркутской государственной сельскохозяйственной академии № от

Рецензенты:

Д.б.н. зав.кафедрой технологии охотничьей продукции и лесного дела

Леонтьев Дмитрий Федорович

Д.б.н. зав.кафедрой прикладной экологии и туризма

Саловаров Виктор Олегович

Предназначено для бакалавров направлений 020400.62 «Биология» и 250100.62 «Лесное дело»

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В процессе любых научных, особенно экспериментальных исследований, так же как и во всех областях практической биологии (медицине, агробиологии, охотоведении, лесоводстве и т.д.), мы имеем дело с цифрами: данными о размерах, весе, возрасте, плодовитости организмов, продуктивности экосистем, урожайности, соотношении между признаками и прочими количественными показателями и их числовыми характеристиками. За кажущимся хаосом этих цифр прячутся конкретные закономерности, которые требуют объективной оценки и научного объяснения. И здесь самое широкое применение находят разнообразные методы и приемы биометрии – вариационной (математической) статистики, призванной с помощью соответствующего математического анализа выразить и оценить разнообразные связи и зависимости между анализируемыми биологическими явлениями.

Биометрия представляет собой своеобразный инструмент, способный выразить в числе и измерить значимость и надежность полученных результатов, заранее рассчитать и спланировать необходимую численность объектов для того или иного эксперимента, оценить достоверность проверяемой в эксперименте гипотезы.

Биометрия представляет собой раздел высшей математики, называемый вариационной статистикой.

Вариационная статистика – это наука, разрабатывающая методы изучения варьирующего признака на массовых материалах в различных областях знаний.

Варьирующими признаками называют признаки, проявляющие определенную закономерность в изменчивости своих значений.

Игнорирование и недооценка статистической обработки и математического анализа полученного материала может свести на нет

результаты многих важных опытов, привести к необоснованным и даже ошибочным заключениям. Напротив, умелое применение биометрических методов увеличивает информационную ценность проведенного исследования, помогает правильно планировать постановку опытов, глубоко разбираться в полученных данных, объективно оценить результаты наблюдений, выявить скрытые закономерности и правильно их трактовать, что в конечном итоге делает биологию точной наукой.

При этом следует иметь в виду, что сама по себе статистическая обработка данных, как бы ни была она совершенна с точки зрения математики, не может служить гарантией качественности выполнения биологом исследования и не способна обеспечить надежность полученных им результатов, если само исследование проведено неправильно или использованные данные ошибочны.

СРЕДНИЕ ВЕЛИЧИНЫ

Для того чтобы получить характеристики не отдельных объектов, а всей группы в целом, определяют среднюю величину признака. В зависимости от исследуемых объектов и от поставленных целей среднюю величину вычисляют различными способами.

Имеется несколько средних величин:

M – средняя арифметическая;

G – Средняя геометрическая;

S – Средняя квадратическая;

H – средняя гармоническая;

Mo – мода;

Me – медиана.

Средняя арифметическая M

Средняя арифметическая (M) – наиболее распространенный и широко применяемый статистический показатель среднего значения варьирующего признака при количественном его выражении.

Она представляет то значение признака, которое имел бы каждый объект, если бы все объекты были одинаковы.

Средняя арифметическая – величина, сумма отрицательных и положительных отклонений от которой равна нулю.

Формула, с помощью которой в простых и наиболее общих случаях выражается значение средней арифметической, выглядит следующим образом:

$$M=\sum V/n,$$

где **M –** средняя арифметическая;

$\sum$ - символ суммирования;

V – результат измерения признака у каждого объекта;

n – число особей в группе или число наблюдений.

Пример: средняя для пяти дат (1,2,3,4,5) равна:

$$\mathbf{M} = \frac{\mathbf{1+2+3+4+5}}{\mathbf{5}} = \mathbf{3}$$

Простая средняя арифметическая – самый распространенный, но не единственный обобщающий показатель для характеристики варьирующих явлений. Часто для практических и научных целей необходимо объединить полученные для однородного материала средние и на этой основе найти общее среднее, характеризующее весь изученный материал с учетом частоты повторяемости варианта. Такую среднюю называют **средней взвешенной.**

Чтобы рассчитать среднюю взвешенную необходимо каждое значение признака (каждую вычисленную среднюю арифметическую) помножить на его «вес» (частоту встречаемости), все эти произведения сложить и сумму разделить на сумму весов.

Взвешенную среднюю арифметическую рассчитывают по следующей формуле:

$$\mathbf{M_{взв.}} = \frac{\sum \mathbf{V_p}}{\sum \mathbf{p}} = \frac{\mathbf{V_1 p_1 + V_1 p_2 + \cdots}}{\mathbf{p_1 + p_2 \cdots}},$$

где **V** – значение признака;

p – математический «вес» усредняемого значения.

Пример:

Необходимо определить среднюю численность грызунов, исходя из данных учетов, согласно которым в ельниках показатель численности составил 4,1, в сосняках – 1,0, в лиственных лесах – 7,2, на болотах – 0,3, в культурном ландшафте – 6,5.

Средняя арифметическая в этом случае будет равна сумме значений признака, деленной на число измерений:

$$\mathbf{M} = \sum \frac{\mathbf{V}}{\mathbf{n}} = \frac{\mathbf{4,1 + 1,0 + 7,2 + 0,3 + 6,5}}{\mathbf{5}} = \frac{\mathbf{19,1}}{\mathbf{5}} = \mathbf{3,82}$$

Однако такой средний показатель неверно отражает реальное состояние численности грызунов на изученной территории (завышая ее более чем вдвое), так как различные биотопы представлены там неодинаково: ельники составляю 10% общей площади, сосняки – 45%, лиственные леса – 7%, болота – 34%, культурные участки – 4%. Гораздо более показательна средняя взвешенная численность:

$$\mathbf{M_{взв}} = \frac{\mathbf{4,1 \cdot 10 + 1,0 \cdot 45 + 7,2 \cdot 7 + 0,3 \cdot 34 + 6,5 \cdot 4}}{\mathbf{100}}$$
$$= \frac{\mathbf{41,0 + 45,0 + 50,4 + 10,2 + 26,0}}{\mathbf{100}} = \frac{\mathbf{172,6}}{\mathbf{100}} = \mathbf{1,73}$$

Приведенные способы вычисления простой и взвешенной средней арифметической удобны в тех случаях, когда число наблюдений (**n**) в выборочной совокупности, подвергнутой обработке меньше 30, и выборка изображается в виде простого статистического ряда, образованного выписанными подряд значениями варьирующего признака.

В биологических материалах чаще всего приходится иметь дело с большим числом наблюдений. Для таких выборок существуют другие приемы вычисления средних, а также и других статистических величин.

При большом числе наблюдений в выборке вычислительной обработке подвергается не простой статистический ряд, а ряд, составленный из классов варьирующего признака, по которым произведена разноска данных, образующих значения частот (**p**).

Обработка большой выборки осуществляется, прежде всего, путем составления вариационного ряда, то есть расположения данных в определенном порядке (ранжировка).

Вариационный ряд – это ряд цифр по величине изучаемого признака, расположенных по возрастающей или убывающей степени с соответствующими им частотами появления признака.

Ступени, на которые разбивается весь вариационный ряд, называют **вариациями** или **классами**.

В практической работе приходится иметь дело с такими признаками, величина которых колеблется в огромных пределах, и отдельные особи отличаются друг от друга на самые разнообразные величины. В этих случаях принято объединять в каждый класс особей с определенной величиной изменчивости, установленной заранее (например: 6-10, 11-15, 16-20 мм и т.д.). Количество классов берется произвольно, при очень точных расчетах – до 15-20, но чаще 8-10.

Порядок составления вариационного ряда:

1. Найти лимит (**lim**), то есть минимальную и максимальную величины, и определить размах колебаний изучаемого признака (промежуток между **min**. и **max**., путем вычитания **min**. от **max.**);
2. Определить число классов вариационного ряда;
3. Найти величину интервала **i** (допустимо округление), для чего нужно разделить полученную при вычитании **min.** от **max**. разницу на намеченное число классов;
4. Установить границы классов (начало и конец);
5. Вычислить среднюю величину признака в каждом классе;
6. Произвести разноску материала в соответствующие классы методом конверта.

Пример построения вариационного ряда:

получены следующие данные о весе 63 взрослых обыкновенных бурозубок (г): 9,2 11,6 8,1 9,1 10,1 9,6 9,3 9,7 9,9 9,6 9,9 7,6 10,0 9,7 8,4 8,6 9,0 8,8 8,6 9,3 **11,9** 9,3 9,2 10,2 11,2 8,1 10,3 9,2 9,8 9,9 9,3 9,1 9,4 9,6 **7,3** 8,3 8,8 9,2 8,0 8,6 8,8 9,0 9,5 9,1 8,5 8,8 9,7 11,5 10,5 9,8 10,0 9,4 8,7 10,0 7,9 8,6 8,7 9,1 8,2 9,2 9,4 8,8 9,8.

Находим минимальный и максимальный показатели и убеждаемся, что вес животных колеблется от **7,3** до **11,9** г.

Размах колебаний признака – **lim = 4,6.**

Округляем эту величину до 5 и строим вариационный ряд из 10 классов с интервалом **i=0,5**.

Классы	Средняя величина класса v	Частоты p
7,1-7-5	7,3	1
7,6-8,0	7,8	3
8,1-8,5	8,3	6
8,6-9,0	8,8	13
9,1-9,5	9,3	17 Условно средний класс A= 9,3
9,6-10,0	9,8	15
10,1-10,5	10,3	4
10,6-11,0	10,8	0
11,1-11,5	11,3	2
11,6-12,0	11,8	2
		∑p= 63

Данные вариационного ряда подвергают дальнейшей статистической обработке, как для получения значения средней арифметической (**M**), так и других статистических показателей.

Обработка вариационных рядов может осуществляться приемом, называемым **«Метод условных отклонений с применением способа произведений»**, который может быть использован и при способе сумм.

Формула средней арифметической по способу произведений выражается следующим образом:

$$M = A + K \cdot \frac{\sum pa}{n},$$

где **M** – средняя арифметическая;

A - условная средняя, приближающаяся по своему значению к средней арифметической;

K – величина класса;

p – частоты вариационного ряда;

n = ∑**p** – число наблюдений в данном вариационном ряду или объем данной совокупности;

a – условное отклонение каждого класса от класса, в котором находится условная средняя (**a**), выраженное числом классов.

По этой формуле произведем вычисления по данным вариационного ряда для обыкновенных бурозубок.

После того как написан ряд из значений классов (**V**) и ряд частот (**p**), производят выделения класса, серединой которого является значение условной средней (**A**).

В качестве условного среднего класса рекомендуется брать тот класс, который занимает центральное место и имеет большее число наблюдений, то есть большое значение частот (**p**) по сравнению с другими классами.

Значение **A** представляет собой середину нулевого класса.

Выделив класс с условной средней (**A**), отделим его от остальных и примем за нулевой от которого по порядку нумеруем остальные классы, что и будет выражать условное отклонение (**a**) каждого класса от нулевого.

В сторону уменьшения признака (в нашем примере от нулевого класса вверх) условные отклонения (**a**) для каждого класса будут иметь знак минус (**-a**), а в сторону увеличения признака (вниз от нулевого класса) условное отклонение будет со знаком плюс (**+a**).

Запишем в таблице графу из значений (**a**).

Классы W	Частоты p	Условное отклонение a	pa
7,1-7-5	1	-4	-4
7,6-8,0	3	-3	-9
8,1-8,5	6	-2	-12
8,6-9,0	13	-1	-13
9,1-9,5	17	0	0
9,6-10,0	15	+1	+15
10,1-10,5	4	+2	+8
10,6-11,0	0	+3	0
11,1-11,5	2	+4	+8
11,6-12,0	2	+5	+10
	∑p=63		∑pa=+3

Согласно формуле средней арифметической, требуется найти для каждого класса произведение **pa**, что и записывается в четвертом столбике.

После этого находим сумму **pa**.

Подставим полученные значения в формулу **M**:

$$\mathbf{M = A + K \cdot \frac{\sum pa}{n} = 9,3 + 0,5 \cdot \frac{3}{63} = 9,32\ (г)}$$

Вычисление средней арифметической для альтернативных признаков

Средней арифметической для альтернативных признаков служит показатель доли, которую составляют члены совокупности, имеющие данный альтернативный признак.

Это можно выразить следующей формулой:

$$\mathbf{M_{альт.} = \frac{p}{n}},$$

где **p –** число членов совокупности с наличием альтернативного

признака;

n – общее число членов выборки.

Например: требуется определить среднее арифметическое число рожденных самцов в пометах кабанов, то есть их долю в общем количестве рожденных поросят. Допустим, что родилось 200 поросят, в том числе 120 хрячков. Средняя арифметическая рождения животных этого пола составит:

$$\mathbf{M_{альт.} = \frac{p}{n} = \frac{120}{200}, или\ 60\%}$$

Свойства средней арифметической

В средней арифметической происходит как бы устранение варьирования признака и установление его обобщающего абстрактного среднего уровня, присущего для данной совокупности при конкретной изменчивости признака. Таким образом, средняя арифметическая является обобщенным статистическим параметром или статистической характеристикой среднего уровня варьирующего признака.

Средняя арифметическая – величина абстрактная, так как при ее вычислении можно получать такие дробные значения, которые в действительности не могут иметь место в связи с природой самого признака.

В тоже время, средняя величина имеет конкретное выражение, показывая величину признака в том же именовании, которым он измерялся.

Основное свойство средней арифметической состоит в том, что сумма отклонений каждого варианта от средней арифметической данной совокупности всегда равна нулю:

$$\mathbf{\sum (V\text{-}M) = 0}$$

То есть если произвести отклонение каждого члена выборки по значению его варьирующего признака от значения средней арифметической, то такая сумма всегда будет равна нулю.

Следующее свойство средних арифметических заключается в том, что сумма отклонений варьирующего признака от любой другой величины, например от условной средней (**A**), неравной средней арифметической, будет всегда больше нуля:

$$\sum (V-A)>0$$

Это выражение используется при вычислении средней арифметической при большом числе наблюдений.

Еще одно свойство средней арифметической можно выразить следующими математическими значениями:

$$\sum(V-M)^2 < \sum(V-A)^2$$

Это означает, что сумма квадратов отклонений от суммы средней арифметической всегда меньше суммы квадратов отклонений, взятых от любого другого числа, отличающегося от средней арифметической.

Кроме указанных основных свойств, средняя арифметическая имеет и другие особенности, которые используются в формулах различных статистических величин.

Средняя геометрическая G

Средняя геометрическая – это величина, которая выявляет средний прирост (или среднее уменьшение) какого-либо показателя за определенный период времени.

Средняя геометрическая необходима для определения среднего значения признака, если он характеризует темп роста, темп увеличения численности популяции. Особенно она удобна в тех случаях, если признак выражен в долях единицы или в процентах и изменяется во времени и по периодам.

Формула средней геометрической:

$$G = \sqrt[n]{v_1 \cdot v_2 \dots v_n},$$

где **G -** средняя геометрическая;

v – значение варьирующего признака;

n – число наблюдений в выборке.

Под корнем стоит произведение вариантов, число которых равно числу наблюдений. Корень имеет степень, соответствующую числу наблюдений.

Для упрощения расчетов в тех случаях, когда степень корня больше двух, производят логарифмирование формулы, а затем по полученному логарифму **G** находят ее абсолютное значение.

Логарифм корня равен логарифму подкорневой величины, деленной на показатель степени корня, а логарифм произведения равен сумме логарифмов, взятых для каждого члена произведения. Следовательно, логарифмирование формулы средней геометрической даст выражение:

$$\lg G = \frac{\lg V_1 + \lg V_2 + \dots + \lg V_n}{n} = \frac{\sum \lg V}{n}$$

Так, если значение варьирующего признака (**V**) при пяти наблюдениях было 5, 8, 10, 12, 13, то значение **G** логарифмированием вычисляется следующим образом:

$$\lg G = \sqrt[5]{5 \cdot 8 \cdot 10 \cdot 12 \cdot 13}$$

$$\lg G = \frac{\lg 5 + \lg 8 + \lg 10 + \lg 12 + \lg 13}{5}$$

$$= \frac{0,6990 + 0,9031 + 1,000 + 1,0792 + 1,1139}{5} = \frac{4,7952}{5}$$

$$= 0,9590$$

По этому логарифму **G** находим (по таблицам логарифмов) абсолютное значение **G**: так как **lgG=0,959**, то **G=9,1**.

Свойств средней геометрической:

1. Произведение чисел ряда, для которого вычисляется средняя геометрическая, всегда равно произведению, полученному от возведения ее в степень, равную числу членов ряда:

$$G^n = V_1 \cdot V_2 \cdot \ldots V_n$$

Это свойство средней геометрической используется для проверки правильности ее вычисления.

В нашем примере это свойство выражается следующими данными:

$$G^5 = 9{,}1 \cdot 9{,}1 \cdot 9{,}1 \cdot 9{,}1 \cdot 9{,}1 = 62403$$

$$G^5 = 5 \cdot 8 \cdot 10 \cdot 12 \cdot 13 = 62400$$

Поскольку оба конечных произведения практически равны, то средняя геометрическая вычислена правильно.

2. Произведение отношений средней геометрической к числам, которые меньше ее, всегда равно произведению отношений чисел ряда, превышающих ее по своему значению, к своей средней геометрической.

По данным нашего примера, это правила должно дать такое равенство:

Числа ряда превышающие **G**

$$\frac{9,1}{5} \cdot \frac{9,1}{8} = \frac{10}{9,1} \cdot \frac{12}{9,1} \cdot \frac{13}{9,1}$$

Числа ряда меньше **G**

или $\frac{82,81}{40} = \frac{1560}{753,571}$, что дает **2,070**

Это свойство средней геометрической наиболее важное и указывает на то, что она представляет собой среднюю из отношений. В связи с этим, служит хорошим показателем для вычисления среднего прироста различных показателей во времени. Поэтому, среднюю геометрическую следует

применять в тех случаях, когда показателями служат отношения, выраженные в долях единицы или в процентах.

Средняя геометрическая особенно пригодна для таких вариационных рядов, у которых имеется ассиметричное распределение частот.

Для определения прироста какого-либо показателя за определенный период времени на основании данных о приросте по частным периодам, составляющим общий период времени пользуются формулой среднего прироста:

$$\mathbf{x = G - 1 = \sqrt[n]{(1 + a_1) \cdot (1 + a_2) \dots (1 + a_n)}}$$

$$\text{или } \mathbf{x = G - 1 = \sqrt[n]{\sum lg(1 + a)},}$$

где **x –** средний прирост за **n** периодов равной длительности;

a - фактический (или плановый) прирост за каждый период, выраженный в долях (%/100);

n - число периодов, за которое определяется прирост;

G - средняя геометрическая.

Для вычисления прироста (**x**) требуется прежде определить среднюю геометрическую (**G**), используя при этом логарифмирование.

Разберем на примере вычисление среднего прироста за общий период по данным процентного прироста за частные и более короткие, но равные периоды.

Пример: определить среднегодовой прирост древостоя за пятилетку, если на каждый год имеется следующее увеличение: за первый год – 5%, за второй – 10, за третий 12, за четвертый – 15, за пятый год – 20%.

Год	Процент	Прирост,	1+a	Lg (1+a)

	прироста по периодам	выраженный в долях (%/100) a		
1	5	0,05	1,05	0,0212
2	10	0,10	1,10	0,0414
3	12	0,12	1,12	0,0492
4	15	0,15	1,15	0,0607
5	20	0,20	1,20	0,0792
				∑lg (1+a) = 0,2517

Произведем расчет необходимых показателей для каждого периода, исходя из формулы прироста:

$$G = \sqrt[n]{\sum lg\,(1+a)}$$

$$lgG = \frac{\sum lg(1+a)}{n} = \frac{0,2517}{5} = 0,0503$$

Откуда: **G = 1,123, x = G-1 + 1,123,** или **= 12,3%**

Таким образом, зная среднюю геометрическую, которая равна 1,123, можно определить средний годовой прирост за пятилетку. Он оказался равным 12,3%

Если же этот прирост был определен через среднюю арифметическую, то он был бы иным, а именно:

M= 5+10+12+15+20/5 = 12,4%, а не **12,3%**, как это получено с применением средней геометрической.

Следует иметь в виду, что чем больше будут различия в показателях прироста, тем больше будет расхождение в показателях среднего прироста, вычисленных с использованием **G** и **M**, и тем более будет ошибочное суждение о приросте, сделанное на основании вычисления по средней арифметической.

Средняя квадратическая S

Средняя квадратическая используется для признаков, которые характеризуются площадью круга и для ее получения измеряют величину диаметра.

К их числу могут быть отнесены диаметры мышечного волокна на поперечном срезе, семенных канальцев половых желез самца, альвеолярных пузырьков молочной или щитовидной желез, колоний микробов, вакуолей у инфузорий, отдельных клеток или их ядер и т д.

Среднюю квадратическую вычисляют по формуле:

$$\mathbf{S} = \frac{\sqrt{\sum \mathbf{V^2}}}{\mathbf{n}},$$

где $\mathbf{V^2}$ - значение варьирующего признака, взятое в квадрате;

n – число наблюдений.

То есть она равна корню квадратному из суммы квадратов признаков, деленному на их число.

Пример: имеется пять дат: 1, 4, 5 ,5, 5 ,средняя квадратическая равна:

$$\mathbf{S} = \sqrt{\frac{1^2 + 4^2 + 5^2 + 5^2 + 5^2}{5}} = \sqrt{\frac{92}{5}} = \sqrt{18,4} = 4,3$$

Средняя гармоническая H

Средняя гармоническая используется при обработке таких совокупностей, для которых применение других средних невозможно. Она необходима для вычисления средних значений, получаемых во времени.

Для этих процессов характерно, что при увеличении одного показателя другой изменяется в обратном направлении, то есть уменьшается. В связи с этим, средняя гармоническая позволяет обрабатывать такие совокупности, у которых значение варьирующего признака находится в обратном соотношении по отношению к суммарному результату.

Эти особенности совокупностей можно показать на следующем примере: чем быстрее бежит лошадь, тем меньше она тратит времени на прохождение пути.

Среднюю гармоническую рассчитывают по формуле:

$$\mathbf{H} = \frac{\mathbf{n}}{\sum \frac{\mathbf{1}}{\mathbf{V}}},$$

где **n** – число периодов времени;

V – величина варьирующего признака.

Пример: для пяти дат (1, 4, 5, 5, 5) средняя гармоническая равна:

$$\mathbf{H}=\frac{\mathbf{5}}{\frac{\mathbf{1}}{\mathbf{1}}+\frac{\mathbf{1}}{\mathbf{4}}+\frac{\mathbf{1}}{\mathbf{5}}+\frac{\mathbf{1}}{\mathbf{5}}+\frac{\mathbf{1}}{\mathbf{5}}}=\frac{\mathbf{5}}{\mathbf{1,85}}=\mathbf{2,70}$$

Среднюю гармоническую применяют при определении изменяющихся скоростей движения.

Пример: почтовые голуби одной станции к месту кормежки летят со скоростью 50 км/час, а в обратном направлении – со скоростью 40 км/час., требуется выяснить среднюю скорость полета для обоих направлений (расстояния равны).

Сделать это можно, рассчитав простую среднюю гармоническую для двух дат 50 и 40:

$$\mathbf{H} = \frac{\mathbf{2}}{\frac{\mathbf{1}}{\mathbf{50}}+\frac{\mathbf{1}}{\mathbf{40}}} = \frac{\mathbf{2}}{\mathbf{0,020+0,025}} = \frac{\mathbf{2}}{\mathbf{0,045}} = \mathbf{44,44\ (км/час.)}$$

Средняя гармоническая всегда меньше, чем средняя арифметическая, мода, медиана и средняя квадратическая.

Мода Мо

Модой, или модальным вариантом, называется наиболее часто встречающиеся значения.

Модальный вариант может быть выражен как качественным, так и количественным признаком.

Например: модальное число сосков у коров – четыре, хотя есть животные с пятью и шестью сосками.

Для количественных признаков модальным будет считаться та величина признака (веса, размера и т.п.), которым будет обладать большее число объектов (членов) генеральной совокупности в случайной выборке.

Величину моды определяют по следующей формуле:

$$\mathbf{Mo} = \mathbf{V_{Mo}} + \mathbf{K}\frac{\mathbf{p_2} - \mathbf{p_1}}{\mathbf{2p_2} - \mathbf{p_1} - \mathbf{p_3}},$$

где $\mathbf{V_{Mo}}$ – начало модального класса;

К – величина класса;

$\mathbf{p_1}$ – частота класса, предшествующая модальному;

$\mathbf{p_2}$ – частота модального класса;

$\mathbf{p_3}$ – частота класса, следующего за модальным.

Вычислим величину моды для примера, приведенного в таблице:

Вариационный ряд по показателю плодовитости полевки

Класс плодовитости	6-7	8-9	10-11	12-13	14-15	К=2
Р	4	14	21	8	3	∑p=50

$$\text{Мо} = 10 + 2 \cdot \frac{21 - 14}{2 \cdot 21 - 14 - 8} = 10 + 2 \cdot \frac{7}{20} = 10 \cdot 2 \cdot 0,35 = 10 + 0,7 = 10,7 \text{ (шт.)}$$

В данном примере модальным классом является класс, имеющий частоту **p = 21**, начало этого класса $\mathbf{V_{Mo}}$ **=10,** $\mathbf{p_1}$ **= 14,** $\mathbf{p_2}$ **=21,** $\mathbf{p_3}$ **=8**.

Величина модального варианта может совпадать или отличаться от значения средней арифметической.

Модальная величина особенно удобна для характеристики качественных признаков, что имеет распространение при изучение генетических особенностей альтернативных признаков. Например, модальными будут доминантные признаки.

Медиана Me

Медианой называется вариант, значение которого делит всю совокупность наблюдений на две равные части. Одна половина объектов совокупности будет иметь значения варьирующего признака меньше, а другая половина объектов больше ее.

Используют показатель медианы чаще для характеристики качественных признаков.

При определении медианы для количественных признаков при малой числе наблюдений члены выборки записывают подряд в возрастающем порядке. Средний член такого ранжированного ряда будет служить ее показателем.

Например: имеем следующий ряд:

Номер в ряду	1	2	3	4	5	6	7	8	9	10
Показатели	5	4	3	3	3	2	2	2	2	1
					Me			**Mo**		

Вычислим **M, Mo** и **Me**

M = ∑V/n = 27/10=2,7, Mo = 2, Me = 3+2/2 = 2,5

Величину медианы в данном ряду вычисляют как среднее из двух вариант, составляющих середину ряда (5-я и 6-я варианты). Если бы ряд имел нечетное число членов, допустим 9 показателей, то среднее значение имел бы номер 5 с величиной признака 3, тогда **Me=3**.

При большом числе наблюдений величину медианы вычисляют по формуле:

$$Me = V_{Me} + K\frac{i_1 - i_2}{p_{Me}},$$

где V_{Me} – начало класса, в котором находится медиана;

К – величина класса;

i_1- число вариантов или сумма накоплений частот, соответствующих половине всех наблюдений (**n/2**);

p_{Me} – частота медианного класса;

i_2 - число вариантов, или сумма накоплений частот по всем классам, предшествующим медианному классу.

Обработка вариационного ряда для вычисления медианы осуществляется приемом, называемым методом «накопленных частот».

Разберем на примере приведенном, в таблице:

Обработка вариационного ряда по количеству плодов брусники на одном побеге методом накопленных частот для вычисления медианы

Классы варьирующего признака	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	16-17	18-19	Итого
Количество побегов	3	5	10	30	20	15	10	10	5	2	n=110

Накопительные частоты	3	8	18	48	68	83	93	103	108	110	-

Ряд накопленных частот составляется путем последовательного сложения частоты каждого последующего класса с частотами предыдущего.

В первом классе нашего ряда частота равна **3**. В следующем классе она составлена из суммирования первого и второго классов **(3+5=8),** накопленные частоты третьего класса получатся из суммирования второго и третьего **(8+10=18)** и т.д. до последнего класса.

Сумма накопленных частот в последнем классе должна быть равна общему числу наблюдений данного ряда **n** или **∑p.**

В нашем примере в последнем классе эта сумма будет равна **110.**

Исходя из ряда накопленных частот, можно найти данные, необходимые для формулы медианы.

Найдем значение $\mathbf{V_{Me}}$, то есть начала класса, в котором находится значение медианы. Таким классом будет класс, в котором накопленные частоты составляют половину всех наблюдений.

Для нашего примера этот класс должен иметь накопительные частоты не меньше величины $\mathbf{i_1}$.

$$\mathbf{i_1 = n/2 = 110/2 = 55}$$

Следовательно, в таблице медианным классом будет класс (пятый слева), имеющий **68** накопительных частот с границами признака **8-9.** $\mathbf{V_{Me}}$, соответствующее нижней границе, или началу медианного класса, будет равно **8**.

Значение $\mathbf{i_2}$ будет равно сумме накопительных частот по классам, предшествующему медианному, и соответствует накопительным частотам, проставленным в четвертом классе, то есть **48.**

В формулу входит значение $\mathbf{P_{Me}}$ (частоты медианного класса), равное **20**.

Подставим в формулу медианы все необходимые данные и вычислим ее значение:

$$\text{Me} = 8 + 2 \cdot \frac{55 - 48}{20} = 8 + 2 \cdot 0,35 = 8,7$$

ПОКАЗАТЕЛИ РАЗНООБРАЗИЯ (ОПРЕДЕЛЕНИЕ СТЕПЕНИ ИЗМЕНЧИВОСТИ ВАРЬИРУЮЩЕГО ПРИЗНАКА)

Первой характеристикой совокупности объектов, вошедших в выборку, как уже известно, служат средние величины. Но этих показателей совершенно недостаточно для суждения о свойствах совокупности по изучаемому признаку, так как всякая группа состоит из неодинаковых особей, отличающихся друг от друга.

Вторым существенным показателем любой совокупности является мера изменчивости (вариабельности) значений данного признака между особями, составляющими совокупность.

Выявление степени изменчивости признаков между членами совокупности, а также установление особенностей в характере распределений особей в вариационном ряду достигается особыми методами, разработанными вариационной статистикой.

Основными показателями вариации служат следующие статистические величины:

Лимиты **– Lim**;

Дисперсия (варианса**) - δ^2**;

Среднее квадратичное отклонение **– δ**;

Нормированное отклонение **- x** или **t**;

Коэффициент изменчивости - C_V или **V**.

Лимиты Lim

Самый простой способ определения изменчивости состоит в сопоставлении максимального и минимального значения варьирующего

признака у членов данной совокупности, то есть $\mathbf{V_{макс}}$ и $\mathbf{V_{мин}}$. Чем больше разница между этими значениями, тем больше вариабельность признака.

Лимиты показывают размах значений и тем самым характеризуют разнообразие признака в группе.

Например: предположим, что на двух массивах черники урожайность с учетных площадок в 1м^2 составила (в г/м^2):

1-й массив: 640 645 650 655 660 $\mathbf{M_1 = 650}$

2-й массив: 610 630 650 670 690 $\mathbf{M_2 = 650}$

Средняя урожайность на обоих массивах одинакова, но на первом разнообразие по этому признаку гораздо меньше, чем на втором.

В приведенном примере

$$\mathbf{Lim_1 = 640\text{-}660}$$

$$\mathbf{Lim_2 = 610\text{-}690}$$

Величину лимита определяют всегда при обработке выборки, несмотря на то, что он является упрощенным показателем изменчивости. Она может быть использована для статистического анализа даже при отсутствие конкретного вариационного ряда. Но лимиты могут служить только грубым показателем изменчивости и не выявляют ее вполне правильно. Могут быть две совокупности, у которых лимит, и средняя арифметическая будут одинаковы, а истинная изменчивость, выраженная более точными методами, окажется различной.

Дисперсия, или варианса δ^2

Дисперсия указывает на степень разнообразия показателя **V** у членов совокупности**.**

Если представить совокупность, состоящую из одновозрастных особей зайцев, то изменчивость этих животных по длине тела может быть выражена

путем сопоставления длины каждого зайца с величиной, характеризующей среднюю арифметическую этого показателя в данной совокупности, то есть изменчивость выражается отклонением **V** от **M**.

Из данных, характеризующих длину тела пяти зайцев, можно составить простой вариационный ряд, который даст следующее значение вариантов и их отклонений от средней арифметической:

Показатели	Зайцы					
	№1	№2	№3	№4	№5	
Варианты ряда V (в см)	45	40	38	35	32	$M=\frac{\sum V}{n}$ $= 190 = 38\ см$
V-M (в см)	45-38=7	40-38=2	38-38=0	35-38=-3	32-38=-6	∑V-M =0
$(V-M)^2$	49	4	0	9	36	$\sum(V-M)^2$ =98

Так как **∑(V-M)** всегда рана нулю, что является свойством средней арифметической, то для определения степени изменчивости у членов совокупности нашего примера необходимо возвести каждое значение (**V-M**) в квадрат, а затем просуммировать и получить величину **$\sum(V-M)^2$**, что и сделано в последней строчке таблицы.

По второму свойству средней арифметической эта величина будет наименьшей по сравнению с любой другой величиной, взятой вместо **M** в этом выражении. Поэтому оно и используется для измерения изменчивости.

Если затем разделить **$\sum(V-M)^2$** на число наблюдений (**n**), то получим величину дисперсии, выражающую изменчивость признака в данной совокупности:

Таким образом, формула дисперсии будет выглядеть следующим образом:

$$\delta^2=\frac{\sum(V-M)^2}{n},$$

где **- δ^2-** дисперсия;

M – средняя арифметическая;

V – результат измерения признака у каждого объекта;

n - число особей в группе или число наблюдений.

Это означает, что дисперсия, измеряющая изменчивость признака, выражает ее через средний квадрат отклонения каждого члена совокупности от средней арифметической данного признака.

В нашем примере:

Дисперсия имеет большое значение в работах по углубленному анализу изменчивости признака с помощью статистического метода «Дисперсионный анализ».

Среднее квадратичное отклонение δ

Среднее квадратичное отклонение служит основным способом измерения изменчивости. Оно может быть получено из значения дисперсии, если из нее извлечь квадратный корень.

Формула среднего квадратичного отклонения:

$$\delta=\sqrt{\frac{\sum(V-M)^2}{n}}$$

При малом числе наблюдений (**n<30**), эта формула несколько изменяется в знаменателе и приобретает следующий вид:

$$\delta = \sqrt{\frac{\sum (V-M)^2}{n-1}}$$

Среднее квадратичное отклонение – наиболее распространенная статистическая величина для измерения изменчивости как количественных, так и качественных признаков членов совокупности. Показывает, насколько в среднем каждый вариант отклоняется от средней арифметической, вычисленной для данной совокупности. Чем больше значение **δ**, тем больше изменчивость данного признака в совокупности.

Среднее квадратичное отклонение **δ** – величина именованная, она имеет то же именование, что и единица измерения изучаемого признака (кг, см, шт. и т.п.).

В так называемых нормальных вариационных рядах весь размах изменчивости, ограниченный максимальным и минимальным значением варьирующего признака, заключает в себе шестикратную величину среднего квадратичного отклонения.

При этом максимальный вариант отстоит от средней арифметической на значение **+3δ**, а минимальный вариант – на значение **-3δ**.

Поэтому принято весь размах изменчивости выражать такой записью: **$M \pm 3\delta$.**

На основании этой особенности изменчивости в нормальных рядах можно осуществлять некоторые расчеты.

Например, по показателю средней арифметической и значению **δ** можно рассчитать каковы будут значения $\mathbf{V_{макс}}$ и $\mathbf{V_{мин}}$ в данной совокупности.

Пример: средний урожай брусники на изучаемом массиве составил **60 г/м^2**, а изменчивость по учетным площадкам выражается **δ= 5 г/м^2**.

Исходя из этих данных, можно предположить, что наивысшая продуктивность составляет:

$$M+3\delta = 60+3\cdot5= 75 \text{ г/м}^2.$$

Самая низкая урожайность соответственно:

$$M-3\delta = 60-3\cdot5 = 45 \text{ г/м}^2.$$

Следовательно, весь размах изменчивости выражается:

$$V_{мин} = 45 \text{ г/м}^2 \text{ и } V_{макс} = 75 \text{ г/м}^2.$$

По размаху изменчивости можно рассчитать приблизительное значение среднего квадратичного отклонения.

Зная, что в пределах лимита содержится шесть **δ**, вычисляется значение одной **δ**. В нашем примере это даст следующее:

$$\delta = \frac{Lim}{6} = \frac{75-45}{6} = 5 \text{ г/м}^2$$

Такую систему расчета для грубого определения среднего квадратичного отклонения целесообразно делать в тех случаях, когда никаких данных о вариационном ряде нет, а исследователь предполагает возможный размах изменчивости признака, и ему необходимо для ряда статистических вычислений иметь хотя бы ориентировочное значение **δ**.

Вычисление среднего квадратичного отклонения (**δ**) для больших выборок осуществляется путем обработки вариационного ряда, разбитого на классы.

Рабочая формула среднего квадратичного отклонения при большом числе наблюдений:

$$\delta = K\sqrt{\frac{\sum pa^2}{n} - (\frac{\sum pa}{n})^2}$$

Для получения величин, входящих в эту формулу, используется обработка вариационного ряда методом произведений, которых уже был использован при вычислении средней арифметической.

Разберем технику вычисления **δ** методом произведений на том примере, которых был использован для определения веса 63 взрослых обыкновенных бурозубок (г):

Приведем данные таблицы, в которой имелись столбцы со значениями классов (**V**), частот (**p**), условных отклонений (**a**), произведения частот на соответствующее условное отклонение (**pa**). Дополним обработку данных этой таблицы введением столбца, в котором проставлены для каждого класса значения **pa²**.

Чтобы получить для каждого класса значения **pa²**, следует уже имеющиеся значения **pa** умножить на условное отклонение **a** этого класса.

Классы W	Частоты p	Условное отклонение a	pa	pa²
7,1-7-5	1	-4	-4	16
7,6-8,0	3	-3	-9	27
8,1-8,5	6	-2	-12	24
8,6-9,0	13	-1	-13	13
9,1-9,5	17	0	0	0
9,6-10,0	15	+1	+15	15
10,1-10,5	4	+2	+8	16
10,6-11,0	0	+3	0	0
11,1-11,5	2	+4	+8	32
11,6-12,0	2	+5	+10	50
	∑p=63		∑pa=+3	∑pa²= 193

Полученные данные подставим в формулу:

$$\delta = K\sqrt{\frac{\sum pa^2}{n} - \left(\frac{\sum pa}{n}\right)^2} = 0{,}5\sqrt{\frac{193}{63} - (\frac{3}{63})^2} = 0{,}5\sqrt{3{,}063 - 0{,}002}$$

$$= 0{,}5\sqrt{3{,}061} = 0{,}5 \cdot 1{,}750 = 0{,}875\ (г)$$

Расчеты показали, что среднее квадратичное отклонение (**δ**), измеряющее изменчивость веса обыкновенных бурозубок равно **0,875 г**, а значение средней арифметической равно:

$$M = A + K \cdot \sum pa / n = 9{,}3 + 0{,}5 \cdot 3/63 = 9{,}32 \text{ г.}$$

Фактический размах изменчивости веса в данном примере равен

$$V_{макс} = 11{,}9 \text{ г}$$

$$V_{мин} = 7{,}3 \text{ г.}$$

Откуда

$$Lim = 11{,}9 - 7{,}3 = 4{,}6 \text{ (г)}$$

Если же определить теоретический размах изменчивости, используя значение вычисленной **δ**, то теоретический лимит даст

$$V_{макс} = M + 3\delta = 9{,}32 + 3 \cdot 0{,}875 = 9{,}32 + 2{,}62 = 11{,}94 \text{ (г)}$$

$$V_{мин} = M - 3\delta = 9{,}32 - 3 \cdot 0{,}875 = 9{,}32 - 2{,}62 = 6{,}70 \text{ (г)}$$

По этим данным лимит равен:

$$Lim = 11{,}94 - 6{,}70 = 5{,}24 \text{ (г)}$$

Такое расхождение между фактическими значения границ изменчивости и размерами лимита вполне закономерно; вычисленное теоретическое значение **δ** правильнее отражает изменчивость, присущую генеральной совокупности.

Среднее квадратичное отклонение – хороший показатель изменчивости признака, но эта величина именованная и зависит не только от степени варьирования, но и от единицы измерения средней арифметической, поэтому по ней можно сравнивать изменчивость лишь одних и тех же показателей, а сопоставлять изменчивость разных признаков нельзя.

Нормированное отклонение (x или t)

Нормированное отклонение – статистический признак, позволяющий определить изменчивость. С его помощью можно выразить в относительных

единицах (долях **δ**) отклонение каждого конкретного члена совокупности от средней арифметической.

Формула нормированного отклонения:

$$x = \frac{V-M}{\delta}$$

Чем больше нормированное отклонение, тем дальше от средней арифметической отстоит величина **V**.

Знак у **x** показывает, в какую сторону от средней арифметической отклоняется данная варианта, то есть будет ли она меньше **(-x)** или больше **(+x),** чем средняя арифметическая.

С помощью нормированного отклонения можно оценить любое полученное значение по отношению к группе в целом, взвесить его и одновременно освободиться от именованных чисел.

Например, если измеренная нами бурозубка имеет длину тела **62 мм** и хвоста **43 мм**, а средние показатели для популяции (или вида) в целом равны соответственно **61** и **37 мм** при $\delta_1 = 6,4$ и $\delta_2 = 2,3$, то искомые нормированные отклонения равны:

$$x_1 = \frac{62-61}{6,4} = +0,16$$

$$x_2 = \frac{43-37}{2,3} = +2,61$$

Данные нормированные отклонения показывают своеобразие изученного экземпляра. При незначительном увеличении длины тела по сравнению со средним уровнем он отличается относительно длинным хвостом, более чем на две с половиной **δ** превышающим средние показатели.

Нормированное отклонение можно использовать для сравнительной оценки индивидов по одному и тому же признаку. Оно помогает определять так называемые «выскакивающие» варианты и решать вопрос о возможности их отбрасывания как артефактов (исключать из дальнейшей обработки).

Решение о выбраковке резко выделяющихся значений (эти отклонения могли возникнуть в результате неточности измерений, ошибок внимания, методических погрешностей и т.д.) принимается на основании нормирования сомнительных вариант по отношению к их средней арифметической. Этой цели служит формула:

$$T = \frac{V-M}{\delta} \geq T_{st},$$

где **T** – критерий выпада;

V – средняя величина для группы, включающей артефакт;

$\mathbf{T_{st}}$ –стандартные значения критерия выпадов.

Пример: имеются варианты 10, 20, 20, 30, 30, 40, 40, 50, 210

$$\mathbf{n = 9,\ M = 50,\ \delta = 61}$$

$$T_{210} = \frac{210-50}{61} = 2{,}6$$

$\mathbf{T_{st}}$ (по табл.) **= 2,4**

Следовательно, показатель 210 может считаться выпадающим и должен быть исключен из обработки.

Коэффициент изменчивости C_V или V

Методы определения степени изменчивости с помощью лимитов и среднего квадратичного отклонения имеют один недостаток: они дают показатель изменчивости признака в именованных величинах, а не в относительных.

Вследствие этого сопоставление разноименных признаков по величине изменчивости с их помощью невозможно.

Коэффициент изменчивости или вариации $\mathbf{C_V}$ показывает изменчивость признака в совокупности в относительных величинах (в процентах).

В связи с этим использовать его целесообразно в тех случаях, когда необходимо сравнить изменчивость разноименных признаков.

Формула коэффициента изменчивости:

$$C_v = \frac{\delta}{M} \cdot 100\%$$

Из формулы видно, что C_v получается путем процентного выражения среднего квадратичного отклонения **δ** от своей средней арифметической.

Так в примере по определению веса 63 бурозубок средний вес составил 9,32 г, а изменчивость, выраженная через **δ** равнялась 0,875 г.

Отсюда коэффициент вариации будет равен:

$$M = 9{,}32 \text{ г}$$

$$\delta = 0{,}875 \text{ г}$$

$$C_v = \frac{0{,}875}{9{,}32} \cdot 100 = 9{,}39\%$$

Чем больше значение коэффициента вариации, тем больше изменчивость признака у членов совокупности.

Коэффициент изменчивости для альтернативных признаков не вычисляется, так как его заменяет значение **δ**, выраженное в процентах.

Генетическая и селекционная науки, используя статистический анализ, выявили разную степень изменчивости основных селекционных признаков. Эти данные накапливаются не только для характеристики видовых особенностей одноименных признаков, но с развитием частной генетики, установлены коэффициенты изменчивости для одних и тех же признаков у разных пород, семейств, линий, экологических рядов, акклиматизированных групп животных.

Применение коэффициента изменчивости в таком многообразном плане значительно расширяет информацию, ценную для селекционной и генетической работы, а так же для решения практических задач, в

углубленном теоретическом анализе особенностей той или иной популяции животных.

Особенности коэффициента изменчивости:

1. Величина коэффициента изменчивости не должна рассматриваться в отрыве от средней арифметической и стандартного отклонения.
   Например: при близких значениях коэффициента изменчивости, полученных на двух выборках, абсолютные величины **M** и **δ** для этих совокупностей могут быть совершенно на разных уровнях.
   Следовательно, одинаковая или близкая величина коэффициента изменчивости для двух выборок еще не означает, что они качественно близки.
2. Одинаковые величины коэффициента изменчивости двух выборок могут быть результатом разных причин, а именно $\mathbf{C_v}$ может увеличиваться за счет повышенного числителя, то есть более высокой величины **δ** одной выборки, или за счет уменьшенного знаменателя, то есть **M**. Эти особенности необходимо учитывать чтобы не допустить ошибку и не сделать неправильные выводы, беря величину $\mathbf{C_v}$ вне связи с величинами **M** и **δ**.
3. Коэффициент изменчивости целесообразно примнять при изучении динамических рядов. Например, в исследованиях, когда изучают показатели, характеризующие возрастные особенности животных, использование $\mathbf{C_v}$ совместно с **M** и **δ** дает ясное представление о динамике и закономерностях онтогенеза по тому или иному признаку.
4. Коэффициент изменчивости имеет большое значение при планировании объема опыта, так как правильное его установление позволяет получать достоверные статистические параметры.

ТИПЫ ВАРИАЦИОННЫХ РЯДОВ И ИХ ГРАФИЧЕСКОЕ ИЗОБРАЖЕНИЕ

Распределение частот по классам варьирующего признака имеет несколько основных типов, для каждого из которых характерны свои особенности и закономерности.

Основные типы распределения:

- Нормальное
- Биноминальное
- Ассиметричное
- Эксцессивное
- Трансгрессивное
- Пуассона

Выше уже разбирался вопрос о способах составления вариационных рядов, служащих для выявления вариабельности признака у членов случайной выборки.

Выборки имеют ограниченное, хотя и большое число наблюдений и служат способом изучения свойств генеральной совокупности.

Вариационные ряды подразделяются на эмпирические и теоретические.

Эмпирическими являются ряды, составленные на основании конкретных данных, полученных из выборочной совокупности. Количество членов в эмпирических ряда может быть малым и большим, но не достигающим бесконечно большого числа (то есть ∞).

В эмпирических вариационных рядах распределение членов выборки по классам выражается частотами (**p**) или частостями (**p'**).

Частость представляет собой дробь, которая выражается в долях единицы встречаемости членов выборки в том или ином классе

вариационного ряда. Вычисляют ее путем деления частот класса на общий объем выборки.

Формула частости:

$$p' = \frac{p}{n}$$

Сумма частостей по всем классам ряда составляет единицу, то есть $\sum p' = 1$.

Теоретические вариационные ряды отражают закономерности распределения членов совокупности по классам варьирующего признака при бесконечно большом числе наблюдений (то есть $n \to \infty$).

Теоретический вариационный ряд служит пределом, к которому стремится эмпирический ряд при увеличение объема выборки до бесконечности.

В теоретических рядах встречаемость членов совокупности, отклоняющихся на определенную величину от средней арифметической ряда, выражается вероятностью (**P**), а не частотой или частостью. При увеличении числа наблюдений частость стремиться к вероятности. Следовательно, вероятность служит мерой возможности появления объектов с данным отклонением от средней арифметической или мерой возможности появления какого-либо события.

Например, возможность появления мутантов в популяции выражается вероятностью. Рождение в пометах самцов или самок также имеет определенную вероятность.

Вероятность представляет собой дробь, величина которой находится в границах от 0 до 1. Если вероятность события равна 0, то осуществление этого события не произойдет. Если вероятность события равна 1, то это событие с необходимостью будет осуществлено. Если вероятность события больше 0,5, то осуществление его более вероятно, чем неосуществление. При

вероятности события меньше 0,5, его называют маловероятным. Сумма вероятностей двух противоположных событий равна единице.

Пример: если на 1000 оленят рождается 550 особей женского пола, то вероятность составляет

$$P = \frac{550}{1000} = 0,55$$

А вероятность противоположного события, то есть рождения особей мужского пола составляет:

$$\mathbf{Q = 1 – P = 1\text{-}0,55 = 0,45}$$

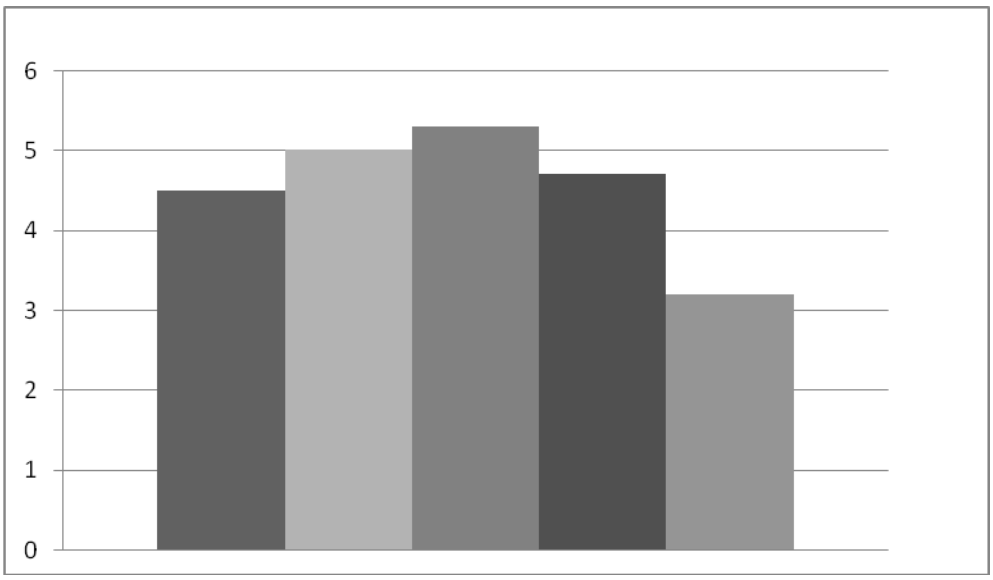
Техника изображения вариационных рядов

Каждый вариационный ряд, составленный достаточно большим числом наблюдений может быть изображен в виде графика. Для этого строят оси координат, в границах которых и будет размещаться кривая, изображающая вариационный ряд.

Вариационная эмпирическая кривая может быть изображена в виде диаграммы, называемой гистограммой, или в виде ломанной кривой, получившей название полигона распределения. Иногда вариационный ряд изображают в виде комуляры.

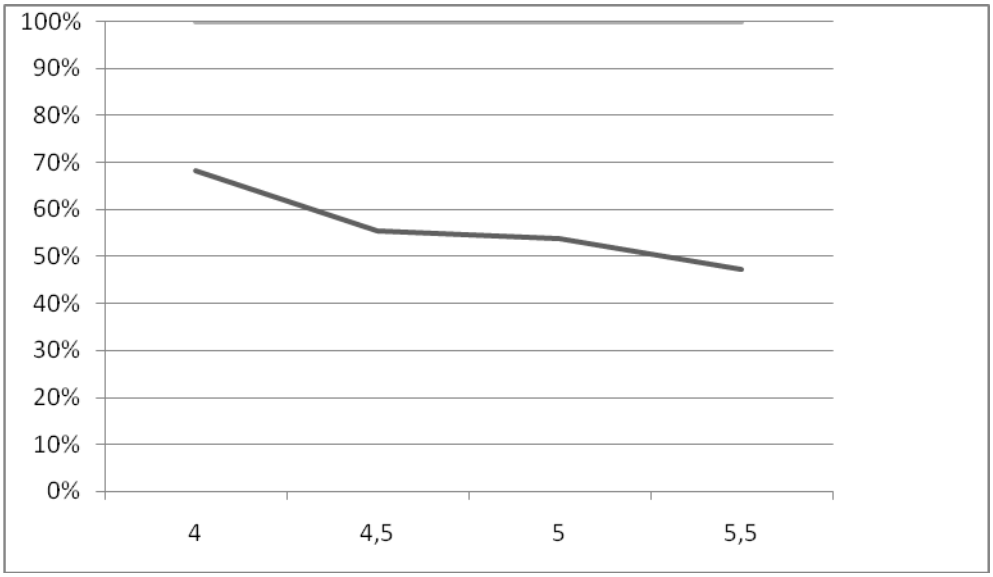
Гистограмму строят в осях координат: на оси **x** откладывают классы варьирующего признака, на оси **y**, откладывают значения частот или их процентное выражение от общего числа наблюдений.

В гистограмме над каждым классом вычерчивают столбик, высота которого соответствует частотом или проценту частот класса отложенным по масштабу оси **y**.



Гистограмма

Каждая гистограмма может быть превращена в так называемый полигон распределения, или эмпирическую вариационную кривую. Для этого следует соединить середины классов и получается ломанная кривая, изображающая вариационный ряд.



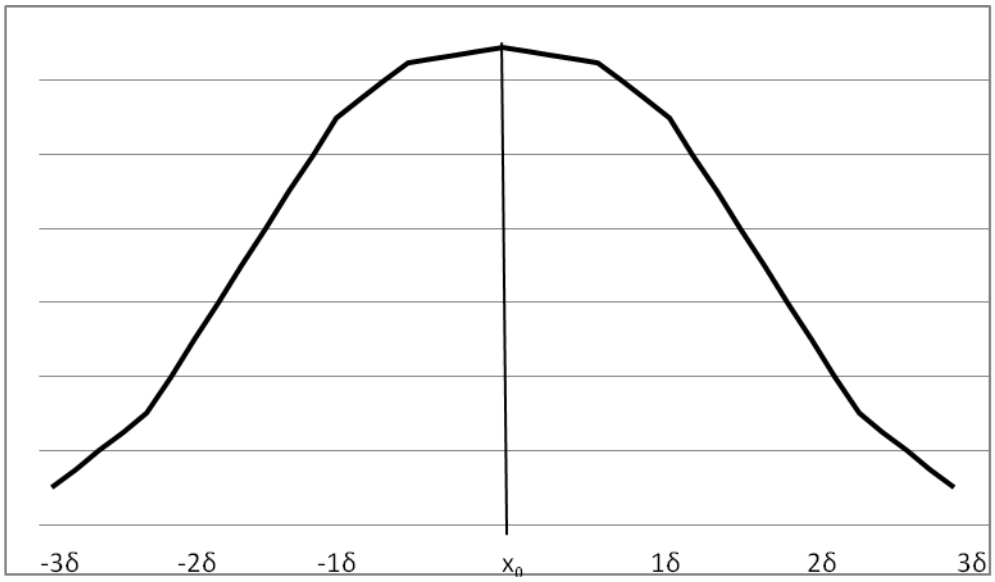
Полигон распределения

Ступенчатость гистограммы и ломанный вид вариационной кривой обусловлены тем, что вариационный ряд имел не большое число наблюдений. Если же число наблюдений будет бесконечным, то есть когда $n \to \infty$, то вариационная кривая приобретает плавный характер и превращается в теоретическую, характеризующую распределение членов генеральной совокупности с теоретическим значением частот. На такой теоретической кривой легче заметить закономерности распределения, его особенности и тип.

По типу распределения частот вариационные ряды довольно сильно могут отличаться между собой.

Нормальное распределение и его свойства

Чаще всего распределение частот у биологических объектов характерезуется нормальным распределением. Если число наблюдений бесконечно большое, то нормальная вариационная кривая имеет следующий вид:



Тип нормальной кривой распределения

Для нормальной кривой характерен ряд закономерностей, проявляющихся в распределении членов совокупности по классам.

Свойства нормальной кривой:

Ось абсцисс (или **x**) служит основанием кривой, на котором откладываются значения варьирующего признака (**V**), выраженные в отклонениях его от средней арифметической в долях **δ**.

Значение **V,** которое соответствует средней арифметической **M**, берут в качестве начальной точки ($\mathbf{x_0}$); вправо от нее откладывают значения варьирующего признака, превышающего **M** и доходящие до $\mathbf{V_{макс}}$, а влево - значения **V** меньше **M,** доходящие до $\mathbf{V_{мин}}$.

Точка $\mathbf{x_0}$ соответствует варианту, величина которого равна **M,** $\mathbf{M_o}$ и $\mathbf{M_e}$.

Размах варьирования признака от $\mathbf{V_{мин}}$ до $\mathbf{V_{макс}}$ почти полностью ограничен отклонениями **V** от **M** на **+3δ** и **-3δ**.

Ось ординат (или **y)** служит для отложения частот (**p**).

Если число наблюдений в совокупности бесконечно большое ($\mathbf{n \to \infty}$), то значения частот (**p**) превращаются в значения вероятности (**P**).

Высота перпендикуляров, восстановленных из значений **V**, будет определять характер плавности боковых границ нормальной кривой.

Перпендикуляр, восстановленный из варианта $\mathbf{V=x_0=M}$ образует вершину нормальной кривой и соответствует максимальной ординате, обозначаемой $\mathbf{y_0}$, которая выражает максимальную частоту и максимальную вероятность для признака, соответствующего средней арифметической, моде и медиане данной совокупности.

Если соединить вершины перпендикуляров, восстановленных из каждого значения варьирующего признака в границах $\mathbf{V_{макс}}$ до $\mathbf{V_{мин}}$, то образуется линия, показывающя нормальную кривую распределения.

Форма нормальной кривой колоколообразная, плавная. Площадь, ограниченная осью **x** и границей линии, образующей кривую, имеет симметричную форму. Если перегнуть рисунок кривой по ординате $\mathbf{y_0}$, то правая и левая части совпадут.

Степень крутизны или округлости боковых ветвей, а так же высота вершины зависят от величины **δ**. Поэтому форма различных вариационных кривых неодинакова.

При одном и том же числе наблюдений, в завистмости от величины **δ**, форма нормальной кривой будет разная:

- чем больше значение **δ**, тем больше будет ее основание и ниже вершина;

- при уменьшении **δ** основание кривой уменьшается, а вершина (высота) кривой увеличивается.

Площадь, ограниченная нормальной кривой, принимается за 1 или за 100%. Она соответствует общему числу наблюдений выборки.

Величина площади, отсеченной ординатами из точек $\mathbf{V_{мин}}$ и $\mathbf{V_{мак}}$, соответствующих **- 3δ** и **+3δ**, составляет **99,7%** от всей площади, то есть от всех наблюдений, вошедших в совокупность.

Площадь, заключенная между ординатой $\mathbf{y_0}$ (ордината точки **M**) и какой-либо ординатой, является определенной величиной и может быть найдена по специальным таблицам (таблицы интеграла вероятности).

Следовательно, если мы знаем, что значение признака отклоняется например от **M** на **+2δ**, то по таблицам интеграла вероятности можно определить, какое число объектов в нормальном вариационном ряду будет иметь величину признака в границах от **M** до **+2δ.**

Пример: средний рост 100 студентов равен 175 см, среднее квадратичное отклонение δ = 15 см. Определить, какое число студентов будет иметь рост на уровне от 175 до 180 см.

Площади и ординаты нормальной кривой распределения (из Миллса)

Нормированное отклонение $x = \frac{V-M}{\delta}$	Вторая функция нормированного оклонения $\varphi(x)$		Первая функция нормированного отклонения f (x) ордината y при значениях $x=\frac{V-M}{\delta}$ То есть вероятность y_x при отклонениях V от M на x
	Площадь между ординатами y_o и y $\frac{V-M}{\delta}$	% числа наблюдений, заключенных между ординатами y_0 и y	
0,0	0,00000	0	0,39894
0,1	0,03983	3,983	0,39695
0,2	0,07926	7,926	0,39104
0,3	0,11791	11,791	0,38139
0,4	0,15542	15,542	0,36827
0,5	0,19146	19,146	0,35207
0,6	0,22575	22,575	0,33322
0,7	0,25804	25,804	0,31225
0,8	0,28814	28,814	0,28969
0,9	0,31594	31,594	0,26609
1,0	0,34134	34,134	0,24197
1,1	0,36433	36,433	0,21785
1,2	0,38493	38,493	0,19419
1,3	0,40320	40,320	0,17137
1,4	0,41924	41,924	0,14973
1,5	0,43319	43,319	0,12952
1,6	0,44520	44,520	0,11092
1,7	0,45543	45,543	0,09405
1,8	0,46407	46,407	0,07895
1,9	0,47128	47,128	0,06562
2,0	0,47725	47,725	0,05399
2,1	0,48214	48,214	0,43398
2,2	0,48610	48,610	0,03547
2,3	0,48928	48,928	0,02833
2,4	0,49180	49,180	0,02239
2,5	0,49379	49,379	0,01753
2,6	0,49534	49,534	0,01358
2,7	0,49653	49,653	0,01042
2,8	0,49744	49,744	0,00792
2,9	0,49813	49,813	0,00595
3,0	0,49865	49,865	0,00443
3,5	0,49977	49,977	0,00087
3,99	0,49997	49,997	0,00014

Для этого находим нормированное отклонение:

$$\mathbf{x}=\frac{\mathbf{V\text{-}M}}{\boldsymbol{\delta}}=\frac{\mathbf{180-175}}{\mathbf{15}}=\mathbf{0,3}$$

То есть рост таких студентов которые превышаеют среднюю на **+0,3δ**.

Ищем по таблице это значение нормированного отклонения, а во второй графе показатель площади, ограниченной ординатами $\mathbf{y_0}$, в соответствующей строке узнаем, что доля площади в этих границах равна **0,11791** от всей площади, принятой за 1, а это соответствует **11,791%** от всех наблюдений, вошедших в совокупность. Следовательно **12** студентов будут иметь рост в границах от **175** до **180** см.

Нормальная кривая характеризуется тем, что любая ее ордината (**y**), то есть значение частот соответствующих любому отклонению варианта от средней арифметичекой **(V-M)**, может быть определена по уровнению нормальной кривой, которое выглядит следующим образом:

$$y_V=\frac{n}{\delta\sqrt{2\pi}}\cdot e^{-\frac{(V-M)}{2\delta^2}},$$

где $\mathbf{y_v}$ – ордината, соответствующая искомой теоретической частоте **p** конкретного значения варьирующего признака;

M – средняя арифметическая;

n - число наблюдений в выборке;

δ – среднее квадратичное отклонение;

e – основание натуральных логарифмов, равное 2,71828;

π- постоянное число, равное 3, 1416

V – величина варьирующего признака, для которого вычисляется теоретическая частота, или ордината $\mathbf{y_v}$

Это уравнение может быть преобразовано и упрощено, что позволяет быстрее найти теоретические частоты для конкретного эмпирического вариационного ряда.

Преобразование формулы исходит из того, что выражение степени у **e** может быть заменено нормированным отклонением:

$$\mathbf{x} = \frac{\mathbf{V\text{-}M}}{\boldsymbol{\delta}}$$

Выраженное в долях **δ**, а выражение $\frac{\mathbf{n}}{\boldsymbol{\delta}\sqrt{2\pi}}$ можно представить:

$$\frac{\mathbf{n}}{\boldsymbol{\delta}} \cdot \frac{\mathbf{1}}{\sqrt{2\pi}}$$

Здесь:

$$\frac{\mathbf{1}}{\sqrt{2\pi}} = \frac{\mathbf{1}}{\sqrt{2 \cdot 3,1416}} = \mathbf{0,39894}$$

То есть эта величина постоянна для любого вариационного ряда.

Имея в виду указанные преобразования и подставив в уравнение постоянные величины, мы получаем уравнение нормальной кривой в следующем виде:

$$y_V = \frac{n}{\delta} \cdot 0,39894 \cdot 2,71828^{-\frac{x^2}{2}}$$

Так как **x** – это нормированное отклонение $\left(\frac{\mathbf{V-M}}{\boldsymbol{\delta}}\right)$, выраженное в долях **δ**, то подставляя различные значения **δ** (от 0,1 до 4), можно заранее вычислить выражение, обозначаемое **∫(t)**, в которое войдут следующие величины:

$$\int (t) = 0,39894 \cdot 2,71828^{-\frac{x^2}{2}}$$

Величина **∫(t)** – первая функция нормированного отклонения.

Зная выражения **∫(t)**, легко вычислить теоретичесчкие частоты вариантов с различной величиной нормированного отклонения, так как остается только определить отношение фактического числа наблюдений к значению **δ** и умножить эти выражения на величину **∫(t)**, которую находят в таблице.

Таким образом, конечное рабочее уравнение, по которым определяют теоретические частоты нормальной кривой, будет выглядеть следующим образом:

$$y_v = \frac{n}{\delta} \cdot \int (t)$$

В этом уравнении **δ** берут в относительных величинах, то есть без умножения на величину класса **К**. Если выражать **δ** в именованных величинах, то формула будет такой:

$$y_V = \frac{n \cdot K}{\delta} \cdot \int (t)$$

Биноминальное распределение

Рассмотренное выше нормальное распределение характерезует распределение особей по количественным признакам. Но у биологических объектов очень часто встречаются качественные альтернативные признаки, такие как пол (самец или самка), тип наследования (доминантный или рецессивный), состояние здоровья (здоровый или зараженный) и т.п.

При изучении варьирования объектов с альтернативными признаками распределение приобретает иную форму, отличающуюся по форме кривой и по закономерностям от нормального распределения. Оно называется биноминальным распределением.

Биноминальное распределение – это частный случай нормального распределения. Оно отражает распределение членов совокупности, имеющих альтернативные признаки.

При биноминальном распределении и таблица вариационного ряда отличается от таблицы вариационного ряда нормального распределения.

Приведем пример биноминального распределения для пометов зайцев по числу рожденных самцов:

Варьирование **V+ (**число самцов в частных группах)	0	1	2	3	4	5	6	7	8
Число пометов, имеющих **V+** (частоты **p**)	1	2	7	9	10	12	7	1	1

Верхний ряд в таблице представляет возможное варьирование по числу рожденных самцов, если число зайчат в пометах было не более 8. Из 50 пометов в одном не было самцов вообще, в двух пометах - по одному самцу, в семи - по 2 самца, в девяти – по 3 самца , в 10 – по 4, в 12 - по 5, в 7 – по 6, в одном – по 7 самцов и один помет полностью состоял из одних самцов (8 голов).

Таким образом, структура биноминального вариационного ряда своеобразна и требует своих приемов обработки.

Особенности биноминального распределения:

1. При биноминальном распределении имеются только два состояния альтернативного признака:признак присутствует **V+** или признак отсутстует **V-** (например, пол мужской или женский)
2. Вероятность появления признака в данной совокупности для всех ее членов постоянна и выражается через **P**. Вероятность отсутствия данного признака также постоянна и выражается **Q =1-P.** Например, вероятность появления доминантного признака у помесей второго поколения при моногибридном расщеплении равна **0,75**, а вероятность его непоявления, то есть появления рецессивного состояния признака, равна **1-0,75 =0,25**.
3. Биноминальное распределение образуется распределением частных групп по классам альтернативного признака **V+**, в которые входят

члены данной совокупности. Число наблюдений (**k**) в каждой группе должно быть одинаковое. Например, для получения биноминального ряда, помещенного в таблице, взято 50 зайчих (это число частных групп), у которых плодовитость была равна 8 зайчатам (это частное число наблюдений в каждой группе, обозначаемое **k**). Распределение частных групп по классам признака **V+** образует ряд частот **p** (второй ряд таблицы).

4. В биноминальном распределении частоты (**p**) могут быть эмпирическими (полученными из данных конкретного материала) и теоретическими. Теоретические частоты определяют с помощью разложения бинома Ньютона $(a+b)^n$. Коэффициенты разложения бинома составляют теоретические частоты биноминального вариационного ряда по классам альтернативного признака **V+**. Это означает, что коэффициенты покажут, как часто (то есть 1, 2…**n** раз) будут встречаться частные группы с отсутствием признака **V+** или с присутствием этого признака у членов совокупности.

Для получения этих теоретических частот бином Ньютона будет выглядеть так:

$$(\mathbf{P}+\mathbf{Q})^{\mathbf{k}},$$

где **P** - вероятность появления альтернативного признака **V+;**

Q - вероятность его отсутствия;

k - число наблюдений в частных группах выборки.

5. Биноминальные ряды характеризуются прерывистостью признака, поэтому кривая бюиноминального ряда имеет ломаную линию.

Форма кривой зависит от величины вероятности **P** и величины **k.**

Если вероятности альтернативного признака равны **(P=Q)**, то есть **P=0,5** и **Q=0,5**, биноминальный ряд симметричен.

Если **P** и **Q** не равны, то ряд будет скошен (ассиметричен). Но если даже **P** и **Q** не равны, а **k** (число наблюдений в частных группах) увеличивается, то ассиметрия биноминального распределения уменьшается, и оно приближается к нормальному распределению.

6. Статистическими характеристиками биноминального распределения служат средняя арифметическая (**M**) и среднее квадратичное отклонение (**δ**).

Если известна вероятность **P** появления признака **V+**, то тогда формулы **M** и **δ** будут следующие:

$$\mathbf{M = k \cdot P}$$

Численность частных групп умножают на вероятность признака **V+**.

$$\delta = \sqrt{\mathbf{k \cdot P \cdot Q}}$$

Численность частных групп **k** умножают на вероятность **P** присутствия **V+** и вероятность **Q** отсутствия **V+**.

Разберем биноминальный ряд распределения помесей кроликов второго покаления по типу шерстяного покрова, если в выборку вошли сведения о 100 окролах и в это число были включены только такие, в которых число крольчат было равно **8 (k=8)**.

Известно, что у помесей второго поколения в связи с расщеплением признаков по правилу Менделя 75% кроликов будут иметь доминантный признак по типу шерсти (нормальношерстность) и 25% помесей не будут иметь доминантной нормальношерстность, а характеризуются наличием рецессивного пухового покрова.

Следовательно, вероятность **P,** появления нормальношерстных кроликов равно **75%**, а вероятность их отсутствия (появления пуховых кроликов) равна **Q=1-P=0,25**.

Имея значения **k, P** и **Q** можно определить среднюю арифметическую и среднее квадратичное отклонение.

$$\mathbf{M = k \cdot P = 8 \cdot 0,75 = 6\ голов}$$

То есть по плодовитости крольчих в 8 крольчат, в среднем 6 голов молодняка будут иметь нормальный шерстный покров.

$$\boldsymbol{\delta = \sqrt{k \cdot P \cdot Q} = \sqrt{8 \cdot 0,75 \cdot 0,25} = \sqrt{1,5} = 1,22 головы}$$

Для биноминальных рядов можно вычислить вероятность появления признака **0,1,2,3…m** раз и по этим вероятностям найти теоретические частоты (**p**) для каждого класса **V+**, принимающего значения $\mathbf{n_+}$ в виде **0,1,2,3…m**.

Для этого пользуются коэффициентами разложения бинома Ньютона $(a+b)^n$. Вместо **a** и **b** вписывают значения вероятности **P** и **Q** данного альтернативного признака, а в степень бинома вместо **n** ставят значение **k** (число наблюдений в частных группах).

После этого проводят разложение бинома по общеизвестной формуле:

$$\mathbf{(P+Q)^k = P^k + \frac{k}{1}P^{k-1} \cdot Q + \frac{k}{2}P^{k-2} \cdot Q + \frac{k}{3}P^{k-3} \cdot Q + \cdots + \frac{k}{k-1}PQ^{k-1} + Q^k}$$

Если по условиям выборки **k=2,** то разложение бинома будет выглядеть так:

$$\mathbf{(P+Q)^2 = P^2 + \frac{2}{1}\ P^{2-1} \cdot Q + Q^2}$$

Здесь коэффициенты у **P и Q** имеют значения **1,2,1.**

Если **k=3**, то разложение бинома дает следующее:

$$\mathbf{(P+Q)^3 = P^3 + \frac{3P^{3-1}}{1} \cdot Q + \frac{3P^{3-2}}{2} \cdot Q^2 + Q^3}$$

Коэффициенты у **P** и **Q** имеют значения **1,3,3,1.**

Для определения коэффициентов бинома при различных значениях следует пользоваться треугольником Паскаля, который выглядит следующим образом:

Число наблюдений в частной группе	Биноминальные коэффициенты
k=1	1 1
k=2	1 2 1
k=3	1 3 3 1
k=4	1 4 6 4 1
k=5	1 5 10 10 5 1
k=6	1 6 15 20 15 6 1 и т.д.

В треугольнике каждый коэффициент является суммой правого и левого коэффициентов, стоящих в предыдущей строке.

Так, если **k= 3**, то крайний левый коэффициент получается от сложения **0+1** из предыдущей строки, что дает **1**; второй коэффициент этого ряда, равный **3**, есть сумма **1+2** из предыдущего ряда; третий коэффициент этого ряда, равный **3**, есть сумма **2+1** и последний коэффициент этого ряда, равный **1**, есть сумма **1+0** предыдущего ряда.

Разберем на примере, как определяют вероятности появления признака $\mathbf{V_A}$ **0,1,2…m** раз и определим теоретические частоты **P.**

10 групп мышей (по 3 мыши в каждой группе) облучали рентгеновскими лучами в дозе 200 рентген. Изучалось появление числа больных животных в каждой группе, и был получен эмпирический биноминальный ряд по числу заболевших мышей:

Число больных мышей в каждой группе $\mathbf{V_A}$	0	1	2	3	k=3
Распределение групп по классам	2	5	2	1	$\sum p = 10$

(частоты **(p)**)					

Требуется найти **M** и **δ** этого биноминального распределения и определить его теоретические частоты (**p**).

Находим среднюю арифметическую **M**:

$$\mathbf{M} = \frac{\sum \mathbf{p} \cdot \mathbf{n}}{\sum \mathbf{p}} = \frac{\mathbf{2 \cdot 0 + 5 \cdot 1 + 2 \cdot 2 + 1 \cdot 3}}{\mathbf{2 + 5 + 2 + 1}} = \frac{\mathbf{12}}{\mathbf{10}} = \mathbf{1,2\ головы}$$

M может быть выражена и следующей формулой:

$$\mathbf{M = k \cdot P},$$

Откуда

$$\mathrm{P} = \frac{\mathbf{M}}{\mathbf{k}} = \frac{\mathbf{1,2}}{\mathbf{3}} = \mathbf{0,4}$$

Следовательно

$$\mathbf{Q = 1 - P = 1 - 0,4 = 0,6}$$

$$\boldsymbol{\delta} = \sqrt{\mathbf{k \cdot P \cdot Q}} = \sqrt{\mathbf{3 \cdot 0,4 \cdot 0,6}} = \sqrt{\mathbf{0,72}} = \mathbf{0,85\ головы}$$

Далее определим теоретические частоты для классов с появлением больных животных **0,1,2** и **3** раза.

Для этого проведем разложение бинома, который по исходным данным нашего примера будет выглядеть так:

$$(\mathbf{P + Q})^{\mathbf{k}} = (\mathbf{0,4 + 0,6})^{\mathbf{3}}$$

По треугольнику Паскаля при **k=3** коэффициенты бинома будут **1,3,3,1**.

Следовательно, разложение дает следующее:

$$(\mathbf{P + Q})^{\mathbf{3}} = \mathbf{P^3} + \frac{\mathbf{3P^{3-1}}}{\mathbf{1}} \cdot \mathbf{Q} + \frac{\mathbf{3P^{3-2}}}{\mathbf{2}} \cdot \mathbf{Q^2} + \mathbf{Q^3}$$

$$\begin{aligned}(\mathbf{0,4 + 0,6})^{\mathbf{3}} &= \mathbf{1 \cdot 0,6^3} + \frac{\mathbf{3 \cdot 0,6^{3-1}}}{\mathbf{1}} \cdot \mathbf{0,4} + \frac{\mathbf{3 \cdot 0,6^{3-2}}}{\mathbf{2}} \cdot \mathbf{0,4^2} + \mathbf{1 \cdot 0,4^3} \\ &= \mathbf{0,216 + 3 \cdot 0,36 \cdot 0,4 + 3 \cdot 0,3 \cdot 0,16 + 0,064} \\ &= \mathbf{0,216 + 0,432 + 0,288 + 0,064 = 1}\end{aligned}$$

Полученные значения этого многочлена и служат показателем вероятности появления больных животных.

Распределение Пуассона

Распределение Пуассона относится к тем случаям, когда имеют дело с появлением редких событий при большом числе опытов, то есть когда вероятность появления этого события очень мала. Варьирующий признак при этом распределении принимает только целые значения: 0,1,2,3 и т.д.

Характерной особенностью распределения Пуассона, четко отличающей его от нормального и биноминального распределения, является то, что значение средней арифметической такого ряда (**M**) совпадает с величиной дисперсии δ^2 или очень близко к ней и, следовательно, этот ряд имеет одну статистическую характеристику. Поэтому, если при вычислении получаются близкие значения **M** и δ^2, то это служит основанием считать данный ряд распределением Пуассона.

В биологии многие признаки характеризуются таким распределением: частота многоплодных рождений у одноплодных видов, появление у нормальных популяций экземпляров альбиносных животных, различных уродств, появление различных мутантных форм и т.п.

Следует иметь в виду, что ряды Пуассона имеют ясно выраженную асимметрию.

Асимметричные ряды

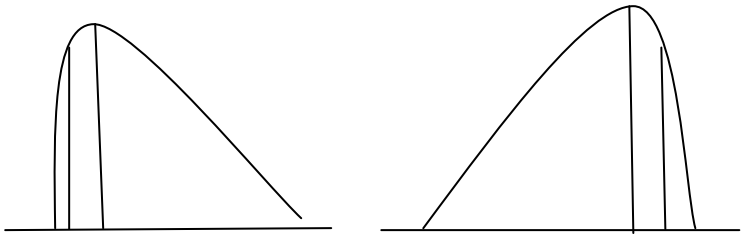
Кроме нормальных, биноминальных и Пуассоновых рядов, при обработке материала могут встречаться скошенные – ассиметричные ряды.

Для асимметричных рядов характерно, что частоты уменьшаются в одну сторону быстрее, чем в другую, а это приводит к смещению вершины кривой в правую или левую сторону от средней арифметической.

У асимметричных рядов и их кривых **M, Mo, Me** не совпадают, как у нормальных кривых.

Если значение **M** лежит правее **Mo,** то такая асимметрия называется положительной или левосторонней**.**

Если **M** лежит левее **Mo,** то такая асимметрия называется отрицательной или правосторонней.



Mo M **M Mo**

Положительная асимметрия Отрицательная асимметрия

Смещение распределения частот в правую или левую сторону от средней арифметической может быть вызвано следующими обстоятельствами:

1. Выборка сделана неправильно: в нее вошло непропорционально мало частот в левой или правой части ряда. Эта причина является методической ошибкой и в работе и не должна допускаться;
2. Сдвиг частот и смещение моды в ту или иную сторону от средней арифметической обусловлен какими-то объективно существующими факторами, которые нарушают обычный для данного признака нормальный характер распределения.

То есть, если причина асимметрии не в методической оплошности, то ее появление будет указывать на происходящие качественные сдвиги в изучаемой совокупности.

Степень асимметрии вариационного ряда определяется с помощью коэффициента асимметрии $\mathbf{A_s}$**:**

$$\mathbf{A_s = \frac{\sum(V - M)^3}{n \cdot \delta^3}},$$

где **V –** значение варьирующего признака;

M – средняя арифметическая;

$\mathbf{(V\text{-}M)^2}$ **–** отклонение от средней в третьей степени;

n - число наблюдений;

$\boldsymbol{\delta^3}$ **–** среднее квадратичное отклонение в третьей степени.

Если $\mathbf{A_s > 0}$, то асимметрия положительная.

Если $\mathbf{A_s < 0}$, то асимметрия отрицательная.

Если $\mathbf{A_s = 0}$, то ряд нормальный.

Если $\mathbf{A_s}$ имеет значение от **0,25** до **0,5**, то это свидетельствует об умеренной асимметрии (косости).

Эксцессивные ряды

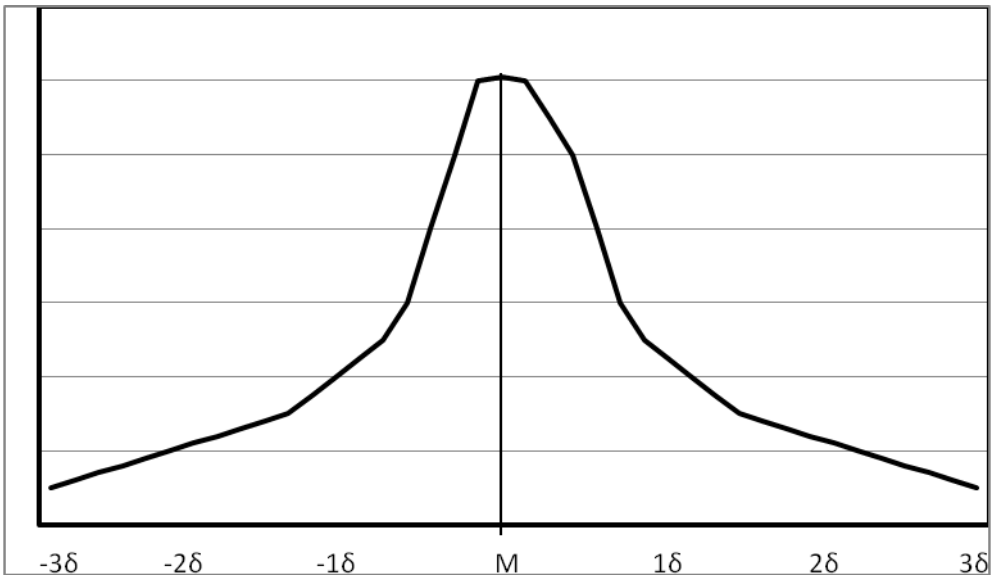
Эксцессивными вариационными рядами называют ряды, у которых значительная доля частот накапливается около варианта, соответствующего средней арифметической.

Модальный вариант при этом имеет значительно большее число наблюдений, чем в нормальных распределениях. Это приводит на графике к высоковершинности и островершинности.

Общий вид такой кривой, называется положительным эксцессом.

Особенность эксцессивных кривых состоит в том, что ее крайние варианты **Vмакс.** и **Vмин.** отстоят от **M** не на **±3δ,** как у нормальных кривых,

а на большее значение **δ** и в приделах **±3δ** находится не 99,7% всех наблюдений, а 97-98%.



Положительный эксцесс

Причины, вызывающие эксцесс, или в неправильно осуществленной выборке или в объективно существующих причинах, уменьшающих частоту появления особей в классах на концах ряда и увеличивающих их накопление в классах, близких к **M**, **Mo** и **Me**.

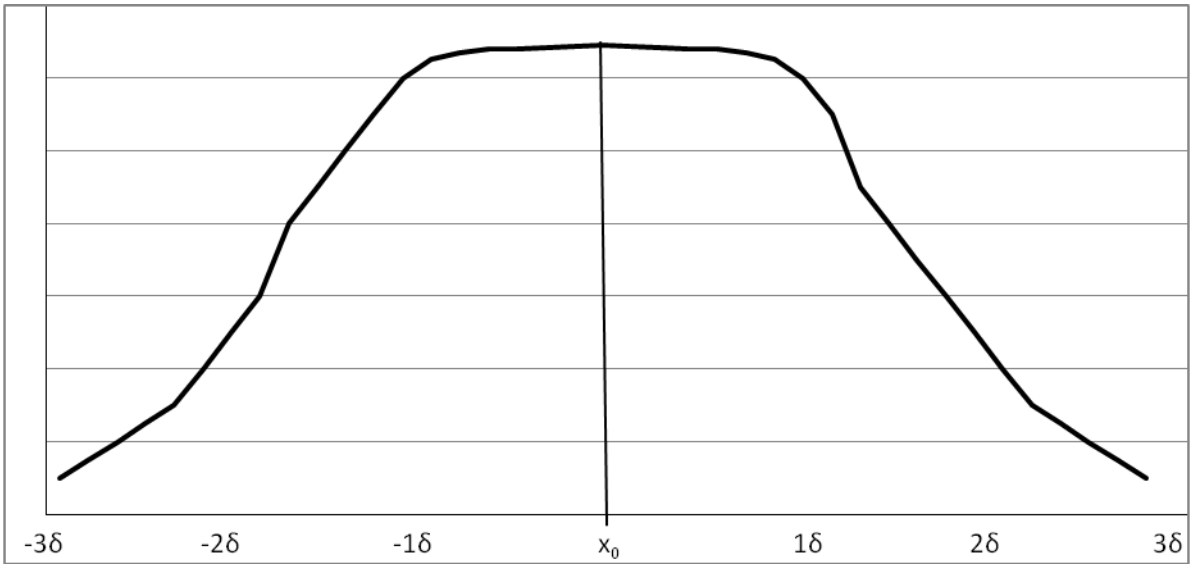
Мерой вытянутости вершины, или мерой эксцесса, служит коэффициент эксцесса **Ex**:

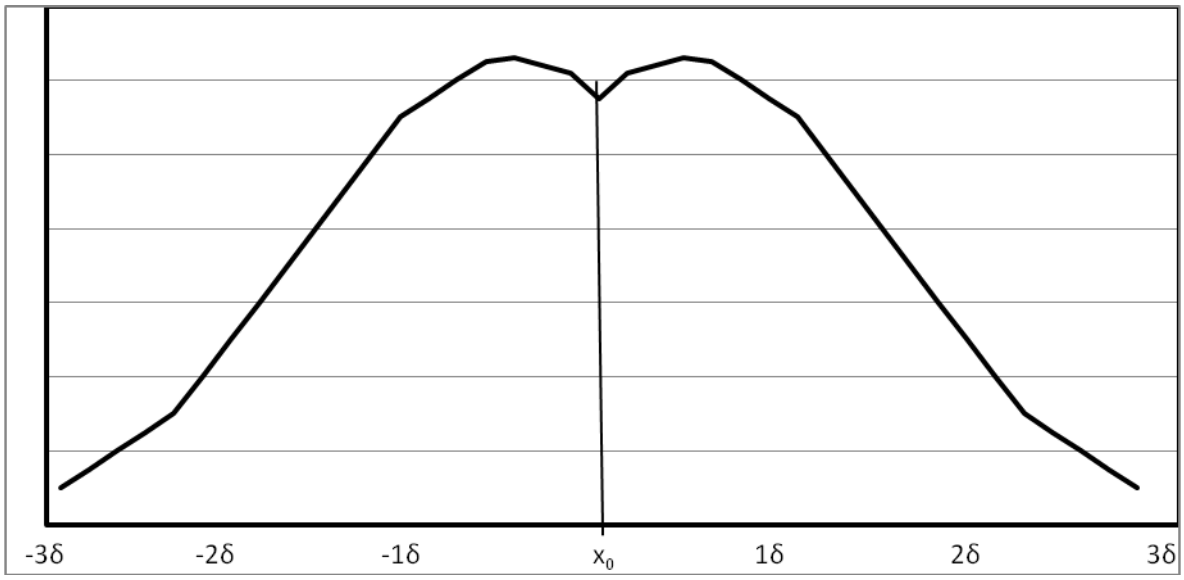
$$\mathbf{Ex} = \frac{\sum(\mathbf{V} - \mathbf{M})^4}{\mathbf{n} \cdot \boldsymbol{\delta}^4} - 3$$

Если эксцесс близок **+0,4**, то это незначительное накопление частот. Если **Ex=0,** то тогда ряд имеет нормальное распределение.

Если эксцесс имеет отрицательный знак, то это свидетельствует о том, что кривая приобретает плосковершинность или двухвершинность.

У плосковершинных кривых ее крайние варианты **Vмакс** и **Vмин** не доходят до границ **+3δ** и **-3δ,** то есть основание плосковершинной кривой меньше, что указывает на сниженную изменчивость признака.



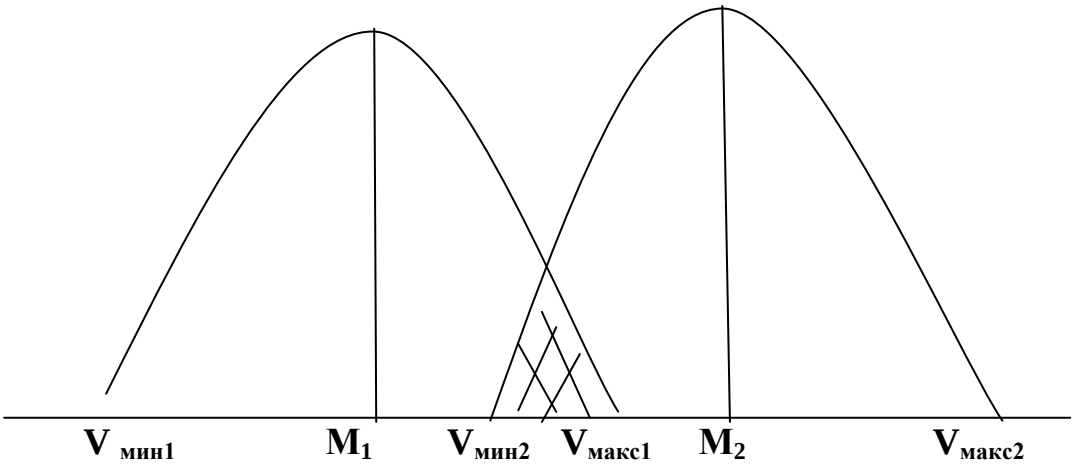


Отрицательный эксцесс

Трансгрессивные ряды и трансгрессивные кривые

Трансгрессивными рядами и кривыми называются ряды, которые отличаются друг от друга величиной средней арифметической и их крайние классы, лежащие около максимального класса первой кривой, служат

минимальными классами другой кривой, что создает в этих частях вариационных кривых их взаимное пересечение.



Общий вид трансгрессирующих кривых

Явление трансгрессии прослеживается при обработке данных различных биологических особей.

При изучении трансгрессивных рядов требуется решить следующие задачи:

1. Определить степень трансгрессии.
2. Определить, достоверна ли разность между средними арифметическими каждого ряда. Если разница между $\mathbf{M_1}$ и $\mathbf{M_2}$ достоверна, то это доказывает наличие двух трансгрессирующих рядов.
3. Определить, к какому из рядов следует отнести конкретную особь, которая имеет признак на уровне вариантов, отчленяющих пересекающиеся части обоих рядов.

Вычислим первый элемент трансгрессирующих рядов.

Определение степени трансгрессии проводится по следующей формуле:

$$\mathbf{T} = \frac{\mathbf{n_1 \cdot p_1 + n_2 \cdot p_2}}{\mathbf{n_1 + n_2}},$$

где **T –** показатель трансгрессии;

$\mathbf{n_1}$ **и** $\mathbf{n_2}$ **–** общее число наблюдений в каждой из выборок;

$\mathbf{p_1}$ **и** $\mathbf{p_2}$ **–** доля трансгрессирующих частот в каждом из рядов, ограниченных площадью кривой между $\mathbf{Vмин_2}$ и $\mathbf{Vмакс_1}$.

Степень трансгрессии может быть большой и малой.

В формуле трансгрессии требуется найти доли трансгрессирующих частот $\mathbf{p_1}$ и $\mathbf{p_2}$. Для этого используется вторая функция нормированного отклонения.

Доли трансгрессирующих частот первого ряда определяются с помощью следующих выражений:

$$\mathbf{p_1 = 0,5 \pm \varphi(x_1)}$$

где $$x_1 = \frac{V_{мин2} - M_1}{\delta_1},$$

где **0,5 –** доля половинной площади от всей нормальной кривой первого ряда;

$\mathbf{\varphi(x_1)}$ **-** буквенное выражение второй функции нормированного отклонения первого ряда (второй столбец таблицы);

$\mathbf{Vмин_2}$ **–** значение минимального варианта второго ряда, который может быть выражен как $\mathbf{Vмин_2 = M_2 - 3\delta_2}$;

$\mathbf{M_1}$ **–** средняя арифметическая первого ряда;

$\mathbf{\delta_1}$ **–** среднее квадратичное отклонение первого ряда.

Доля трансгрессирующих частот второго ряда определяется по аналогичной формуле, только соответственно меняются значки:

$$\mathbf{p_2 = 0,5 \pm \varphi(x_2)}$$

где $$x_2 = \frac{V_{макс1} - M_2}{\delta_2}$$

$$\mathbf{Vмин_2 = M_1 - 3\delta_1}$$

Разберем на примере вычисление величины трансгрессии: имеются два вариационных ряда, характеризующих вес ягод клюквы болотной, собранных в южных и северных районах Иркутской области.

В южных районах: $\mathbf{M_1 = 3,5\ г\ \delta_1 = 0,2\ г\ n_1 = 500}$

В северных районах: $\mathbf{M_2 = 4,2\ г\ \delta_2 = 0,3\ г\ n_2 = 400}$

Вычисляем коэффициенты трансгрессии для первого и второго рядов:

1. Определяем крайние значения вариантов по каждому ряду

   $\mathbf{Vмин_1 = M_1 - 3\delta_1 = 3,5 - 3 \cdot 0,2 = 2,9\ г}$

   $\mathbf{Vмакс_1 = M_1 + 3\delta_1 = 3,5 + 3 \cdot 0,2 = 4,1\ г}$

   $\mathbf{V\ мин_2 = M_2 - 3\delta_2 = 4,2 - 3 \cdot 0,3 = 3,3\ г}$

   $\mathbf{Vмакс_2 = M_2 + 3\delta_2 = 4,2 + 3 \cdot 0,3 = 5,1\ г}$

2. Находим $\mathbf{x_1}$ и $\mathbf{x_2}$ в долях $\boldsymbol{\delta}$, то есть выражаем их через нормированное отклонение

   $$\mathbf{x_1 = \frac{V_{мин2} - M_1}{\delta_1} = \frac{3,3 - 3,5}{0,2} = \frac{-0,2}{0,2} = -1,0}$$

   По таблице находим, что при таком отклонении **V** от **M** значение $\boldsymbol{\varphi(x_1)}$ будет равно **0,34134**

   $$\mathbf{x_2 = \frac{V_{макс1} - M_2}{\delta_2} = \frac{4,1 - 4,2}{0,3} = \frac{-0,1}{0,3} = -0,33}$$

   По таблице находим, что $\boldsymbol{\varphi(x_2)}$ равна **0,11791**

3. Находим площадь кривой, соответствующую доли или проценту частот $\mathbf{p_1}$ и $\mathbf{p_2}$ трансгрессирующих рядов, то есть

   $$\mathbf{p = 0,5 \pm \varphi(x)}$$

   Для первого ряда это выражение берут в виде суммы $\mathbf{0,5+\varphi(x)}$

   $$\mathbf{p_1 = 0,5 + \varphi(x_1) = 0,5 + 0,34134 = 0,84134}$$

   или **84,13%** будут входить в трансгрессирующую часть ряда

   для второго ряда $\mathbf{p_2}$ определяют как разность $\mathbf{0,5 - \varphi(x_2)}$

   $$\mathbf{p_2 = 0,5 - \varphi(x_2) = 0,5 - 0,11791 = 0,38209}$$

или **38,21%** частот второго ряда будут состоять в трансгрессии с частотами первого ряда.

4. Находим коэффициент трансгрессии:

$$T = \frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2} = \frac{500 \cdot 0,84134 + 400 \cdot 0,38209}{500 + 400}$$
$$= \frac{420,67 + 152,836}{900} = \frac{573,5}{900} = 0,6372$$

или **63,72%** частот имеют трансгрессию между обоими рядами.

Второй элемент анализа трансгрессии сводится к определению разности **D** между средними арифметическими каждого ряда:

$$\mathbf{D = M_1 - M_2}$$

Если эта разность будет достоверна (когда статистическая ошибка этой разницы укладывается в ней не менее 3 раз), то такое различие между средними арифметическими обоих рядов будет свидетельствовать о трансгрессивном типе взаимоотношений между рядами.

Если эта разность будет недостоверной, то один ряд как бы является частью другого и суммирование их частот по соответствующим классам даст единую кривую, проявляющую двухвершинность.

Для определения статистической ошибки разности между средними арифметическими обоих рядов используется формула:

$$\mathbf{m_D = \sqrt{m_{M_1}^2 + m_{M_2}^2}},$$

где $\mathbf{m_{M1}}$ **и** $\mathbf{m_{M2}}$ - ошибки средних арифметических каждого ряда.

Формулы этих ошибок:

$$\mathbf{m_M = \frac{\delta}{\sqrt{n}}}$$

Используем данные нашего примера. Найдем статистическую ошибку для $\mathbf{M_1}$ и $\mathbf{M_2}$.

$$\mathbf{m_{M_1} = \frac{\delta_1}{\sqrt{n_1}} = \frac{0,2}{\sqrt{500}} = \frac{0,2}{22,36} = 0,0089 \approx 0,009\ г}$$

$$m_{M_2} = \frac{\delta_2}{\sqrt{n_2}} = \frac{0,3}{\sqrt{400}} = \frac{0,3}{20} = 0,015 \text{ г}$$

Найдем разность средних арифметических:

$$\mathbf{D = M_2 - M_1 = 4,2\text{-}3,5=0,7 \text{ г}}$$

Ошибка разности **D** равна:

$$m_D = \sqrt{m_1^2 + m_2^2} = \sqrt{0,009^2 + 0,015^2} = \sqrt{0,000306} = 0,017 \text{ г}$$

Достоверность разности **D**:

$$\frac{D}{M_D} = \frac{0,7}{0,017} = 41$$

Следовательно, ошибка $\mathbf{M_D}$ содержится **41** раз в разности **D,** что указывает на полную достоверность наличия трансгрессирующего наложения между вариационными рядами.

Третий элемент анализа трансгрессирующих рядов заключается в определении к какому из рядов следует отнести ту или иную особь, у которой величина признака находится в границах вариантов, являющихся общими для обоих рядов, то есть в границах трансгрессирующей части обоих вариационных рядов.

Для определения принадлежности данной особи к тому или иному ряду, образующих трансгрессию, пользуются методом комбинированных признаков (предложенным ихтиологом Гейнеке).

Метод комбинированных признаков основан на сопоставлении суммы квадратов отклонений $(V - M)^2$, вычисленных для трансгрессирующих рядов.

Поясним этот метод следующим примером: имеется ягода клюквы неизвестного происхождения, у которой вес составляет 4 г. Требуется узнать, следует ли ее отнести на основании этого показателя в северной или южной популяции.

Для решения этого вопроса возьмем несколько признаков, которые мало коррелируют друг с другом, но могут служить характеристикой для данных популяций. Положим, что в качестве признаков взяты диаметр плода и содержание витамина С.

Район сбора	Вес плода	Содержание витамина С, мг	Диаметр плода, мм
Южные районы	3,5	500	7
Северные районы	4,2	450	10
Показатели плода А	4,0	470	8,5

Сравним показатели интересующего нас плода **А** со средними показателями плодов собранных в южных и северных районах Иркутской области, выразив это через квадрат отклонения $(\mathbf{V_A} - \mathbf{M_1})^2$ и $(\mathbf{V_A} - \mathbf{M_2})^2$.

Использование метода комбинированных признаков

Признаки	Отклонения показателей плода от показателей плодов южной и северной популяций	
	$V_A - M_1$	$V_A - M_2$
Вес плода (г)	4,0-3,5=0,5	4,0-4,2= -0,2
Содержание витамина С (мг%)	470-500= -30	470-450=20
Диаметр плода (мм)	8,5-7= 1,5	8,5 -10= -1,5
Квадраты отклонений по каждому признаку	$(V_A - M_1)^2$ $0,5^2 = 0,25$ $-30^2 = 900$ $1,5^2 = 2,25$	$(V_A - M_2)^2$ $-0,2^2 = 0,4$ $20^2 = 400$ $-1,5^2 = 2,25$
Сумма квадратов отклонений	$\sum(V_A - M_1)^2$ =902,5	$\sum(V_A - M_2)^2$= 402,65

Следовательно плод **А** по сравниваемым признакам ближе стоит к плодам клюквы, собранным в северных районах Иркутской области, так как:

$$\sum(V_A - M_1)^2 < \sum(V_A - M_2)^2$$

Этот случай удобен в тех случаях, когда исследователь располагает единичными экземплярами, в отношении которых необходимо выяснить, к какой группе особей они ближе относятся.

СТАТИСТИЧЕСКИЕ ОШИБКИ

Статистический метод изучения варьирующего признака основывается на использовании выборочной совокупности особей, взятых по принципу случайности из генеральной совокупности. Следовательно, выборочная совокупность представляет часть, а генеральная совокупность является целым, о котором мы судим по величинам, получаемым при обработке выборочной совокупности.

Выборочный метод, лежащий в основе вариационно-статистической обработки материалов, служит источником статистических ошибок.

Необходимо добиваться того, чтобы ошибки были полностью устранены, а если их нельзя избежать, то следует свести их к минимуму.

Существуют следующие основные типы ошибок в математической обработке экспериментальных материалов:

1. Ошибки, являющиеся следствием просчетов, описок, неверных арифметических вычислений и т.п., происходящие из-за недостаточного внимательного отношения к работе. Устранение их можно осуществить перепроверкой всего исходного материала и более тщательной и внимательной работой.

2. Систематические ошибки, появляющиеся в результате неточного измерения какого-либо показателя. Это происходит из-за неточности или малой разрешающей силы используемых приборов. Так как величина неточности прибора всегда одинакова, то возникает систематическая ошибка измерения. Величину такой ошибки можно учесть, если определить ее путем проведения многократных измерений. Обычно ошибка прибора, то есть его точность, указывается на его паспорте и может быть внесена в виде поправки к проведенному измерению.

3. Статистические ошибки, обусловленные самим статистическим методом. То есть выборочным методом, при котором из генеральной совокупности отбирается по принципу случайности часть ее членов (случайная выборка).

Чем меньше статистическая ошибка, вычисленная по отношению к какой-либо характеристике выборки (например, для **M,δ** и т.д.), тем лучше выборочные данные характеризуют генеральную совокупность.

В статистике разработаны приемы вычисления величины статистических ошибок.

По величине ошибки и соотношению ее с той выборочной характеристикой, для которой она вычислена, можно судить о том, достаточно ли точно выборочные данные отражают характеристики, присущие генеральной совокупности.

Величина статистической ошибки зависит от числа наблюдений **n,** вошедших в выборку, и от степени изменчивости изучаемого признака. Чем больше объем выборки, то есть чем больше в ней число наблюдений, тем меньше будет статистическая ошибка, и в то же время, чем больше изменчивость признака, тем больше статистическая ошибка.

Следовательно, чтобы уменьшить статистическую ошибку, нужно стремиться увеличить объем выборки, особенно в тех случаях, когда изучаемый признак обладает большой вариабельностью.

Статистическая ошибка средней арифметической

Для большой выборки статистическую ошибку принято называть или средней или стандартной. Чаще всего ошибку обозначают буквой **m**, у которой подстрочно записывают значок, указывающий, для какой величины она вычислена. Ошибки имеют то же именование, что и статистические величины, для которых они вычислены.

Для средней арифметической большой выборки ошибка выражается следующими рабочими формулами:

$$m_M = \frac{\delta}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}},$$

где **δ –** выборочное среднее квадратичное отклонение;

n – число наблюдений в выборке;

N – число наблюдений в генеральной совокупности.

Когда **N =∞, то** $\sqrt{1 - \frac{n}{\infty}} = \sqrt{1 - 0}$ и формула ошибки может быть использована в виде

$$m_M = \frac{\delta}{\sqrt{n}}$$

Которой и пользуются, если **n** составляет не более **5-10%** от **N.**

Из этой формулы видно, что ошибка средней арифметической прямо пропорциональна изменчивости признака и обратно пропорциональна числу наблюдений в выборке.

Как пользоваться статистической ошибкой, разберем на примере: из 1000 выкопанных корней радиолы розовой, была взята десятипроцентная случайная выборка, то есть 100 корней, у которых изучали их массу. По данным этой выборки получено следующее: средняя масса корней в воздушно-сухом состоянии **M= 100 г**, среднее квадратичное отклонение **δ = 3,3 г.** Определить, правильно ли отражают эти данные выборки генеральную совокупность, состоящую из 1000 корней.

Для этого определим ошибку средней арифметической:

$$m_M = \frac{\delta}{\sqrt{n}} = \frac{3,3}{\sqrt{100}} = \frac{3,3}{10} = 0,33 \text{ г}$$

Таким образом, ошибка средней арифметической составила 0,33 г.

Обычно записывают статистическую величину совместно с ее ошибкой, а именно:

$$M \pm m_M = 100 \pm 0,33 \text{ (г)}$$

Показателем того, насколько правильно выборочная средняя отражает генеральную среднюю, служит критерий достоверности **t,** который получают путем деления любой средней статистической на ее ошибку.

$$t = \frac{\text{средняя}}{\text{ее ошибка}}$$

В нашем примере:

$$t_M = \frac{M}{m_M} = \frac{100}{0,33} = 303$$

По величине **t** судят о том, насколько правильно выборочная средняя отражает генеральную среднюю. Чем больше значение **t**, тем достовернее выборочная средняя.

Для различных исследований установлены различные уровни **t**, то есть требования к уровню достоверности полученных выборочных величин.

Для производственных и научно-производственных опытов, и большинства биологических работ достаточен критерий достоверности **t=2** или **t=2,5**.Так как при **t=2, P** (величина вероятности)**= 0,95**, то это означает, что в 95 опытах из 100 были получены значения статистических величин такие же, как в том опыте, который проведен и в котором критерий достоверности оказался равен 2 и только в 5 опытах могут быть получены иные характеристики выборочного материала. Следовательно, при **t=2,** что соответствует **P=0,95,** допускается 5%-ная ошибка в достоверности выборочных данных, а это значит, что из 20 опытов один будет неверным.

Если **t= 2,5,** то этому соответствует **P=0,987,** или один неверный опыт на 76 правильных. В работах, где достоверность полученных данных требует

более высокой вероятности, берут критерий достоверности на уровне **t=3,** когда **P=0,997,** что дает один неверный результат и 332 верных опыта.

В тех случаях, где требуется установить дозы лечебных препаратов, дозы безвредного облучения и т.п., то есть в тех случаях, когда выводы от обработки материалов очень ответственны, критерий достоверности **t** берут на уровне **4**, когда **P=999936,** что дает один неверный опыт на 15625 верных.

Ошибка средней арифметической при малом числе наблюдений

При малом числе наблюдений в выборке ошибка средней арифметической вычисляется по формуле:

$$\mathbf{m} = \frac{\boldsymbol{\delta}}{\sqrt{\mathbf{n-1}}},$$

где **δ –** среднее квадратичное отклонение;

n – число наблюдений в выборке;

n-1 =v – число степеней свободы.

Обычное представление о достоверности средней арифметической по показателю критерия достоверности при малой выборке несколько изменяется.

Как показали работы Стьюдента, величина **t** при малом числе наблюдений зависит от величины **n** и поэтому достоверная величина определяется с учетом числа степеней свободы **v**, равным **n-1** (числу наблюдений в выборке, уменьшенному на 1).

Достоверное значение **M** определяется по специальной таблице Стьюдента, в которой приведены значения **v** и **t** при разных уровнях вероятности **P.**

Вычисление границ доверительных интервалов для генеральной средней арифметической при малом числе наблюдений осуществляется с помощью доверительных значений **t**.

Значение критериев достоверности **t** при различных уровнях вероятности **P** и числа степеней свободы **v**, дающие достоверную величину средней арифметической и достоверность разности **($M_1 - M_2$)** при малом и большом числе наблюдений **n**

Число степеней свободы v	Уровень вероятности P		
	0,95	0,99	0,999
	Значение t		
1	12,71	63,7	637,0
2	4,3	9,9	31,6
3	3,2	5,8	12,9
4	2,88	4,6	8,6
5	2,6	4,0	6,9
6	2,4	3,7	6,0
7	2,4	3,5	5,3
8	2,3	3,4	5,0
9	2,3	3,3	4,8
10	2,2	3,2	4,6
11	2,2	3,1	4,4
12	2,2	3,1	4,3
13	2,2	3,0	4,1
14	2,15	3,0	4,1
15	2,1	3,0	4,1
16	2,1	2,9	4,0
17	2,1	2,9	4,0
18	2,1	2,9	3,9
19-20	2,1	2,9	3,9
21-24	2,1	2,8	3,8
25-28	2,1	2,8	3,7
29-31	2,0	2,8	3,7

32-34	2,0	2,7	3,7
35-42	2,0	2,7	3,6
43-62	2,0	2,7	3,5
63-175	2,0	2,6	3,4
176 и больше	2,0	2,6	3,3

Пример: определить границы доверительного интервала урожайности ягод брусники с 20 учетных площадок, если средняя урожайность составила **50 г/м^2**, а среднее квадратичное отклонение **δ=3,5 г/м^2**.

На основании этих данных определяем ошибку $\mathbf{m_M}$.

$$\mathbf{m_M = \frac{\delta}{\sqrt{n-1}} = \frac{3,5}{\sqrt{20-1}} = \frac{3,5}{\sqrt{19}} = \frac{3,5}{4,36} \approx 0,8}$$

Отсюда

$$\mathbf{t = \frac{M}{m_M} = \frac{50}{0,8} = 62,5}$$

По таблице находим, что при числе степеней свободы

$$\mathbf{v = n - 1 = 20 - 1 = 19}$$

При вероятности **P=0,95 t=2,1**

при **P=0,99 t=2,9**

при **P=0,999 t=3,9**

Вычисленное же значение **t** больше табличных, следовательно, полученная на 20 площадках урожайность брусники правильно отражает среднюю урожайность ягод.

Ошибка при альтернативных признаках

При альтернативных признаках выборка образуется особями, подразделяющимися на два класса: группа особей, имеющая данный альтернативный признак, и группа особей, у которой данный альтернативный признак отсутствует.

Для определения ошибки по каждой группе особей пользуются следующей формулой:

$$\mathbf{m_p = m_q = \sqrt{\frac{pq}{n}}},$$

где **p** –число особей, имеющих альтернативные признаки;

q – число особей, не имеющих этого признака;

n – число особей в выборке, равное **p+q.**

Пример: определить ошибку для показателя численности альбиносных и пигментированных мышей, полученных в потомстве второго поколения помесей, если их родители подверглись облучению лучами Рентгена.

В опытной группе 100 мышей второго поколения. **60** голов были пигментированными **(p)** и **40** –альбиносами **(q)**.

Вычислим ошибку для обеих групп (для **p** и **q**).

$$\mathbf{m_p = m_q = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{60 \cdot 40}{100}} = \sqrt{\frac{2400}{100}} = \sqrt{24} = 4,9 \text{ головы}}$$

Величина ошибки для обеих групп всегда одинаковая, то есть

$$\mathbf{m_p = m_q}$$

Но критерий достоверности **t** бывает разным, так как численность групп отличается друг от друга.

Определим $\mathbf{t_p}$ и $\mathbf{t_q}$.

$\mathbf{p \pm m_p = 60 \pm 4,9}$ головы откуда

$$\mathbf{t_p = \frac{60}{4,9} = 12,24}$$

Следовательно, численность пигментированных животных вполне достоверна, так как **t>2**.

$q \pm m_q = 40 \pm 4,9$ головы, откуда

$$t_q = \frac{40}{4,9} = 8,16$$

Численность группы **q** также достоверна.

Ошибка выборочной доли и метод

Ошибку при альтернативных признаках можно получать и в относительных величинах, если численность особей в альтернативных классах выражается в долях единицы, то есть:

$$p' = \frac{p}{n} \text{ и } q' = \frac{q}{n}$$

При этом $p' = 1-q'$ и $q' = 1-q'$

Формула ошибки доли аналогична той, какая употреблялась и при абсолютных величинах численности альтернативных вариантов:

$$m_{p'} = m_{q'} = \sqrt{\frac{p' \cdot q'}{n}},$$

где **n-** число наблюдений, или **n=p+q.**

Для получения более точных значений ошибок доли в знаменателе берут **n-1**, что особенно имеет большое значение при малых выборках.

В нашем примере доля пигментированных мышей составляет:

$$p' = \frac{p}{n} = \frac{60}{100} = 0,6$$

А альбиносных:

$$q' = \frac{q}{n} = \frac{40}{100} = 0,4$$

Ошибки долей будут равны:

$$m_{p'} = m_{q'} = \sqrt{\frac{p' \cdot q'}{n}} = \sqrt{\frac{0,6 \cdot 0,4}{100}} = \sqrt{\frac{0,24}{100}} = \sqrt{0,0024} = 0,049$$

Следовательно, достоверность долей будет следующей:

По пигментированной группе:

$$p' \pm m_{p'} = 0,6 \pm 0,049$$

$$t_{p'} = \frac{p'}{m_{p'}} = \frac{0,6}{0,049} = 12,24$$

По альбиносной группе:

$$q' \pm m_{q'} = 0,4 \pm 0,049$$

$$t_{q'} = \frac{q'}{m_{q'}} = \frac{0,4}{0,049} = 8,16$$

Что указывает на достоверность обеих долей.

Если численность особей в каждой группе альтернативных признаков выражается в процентах, то ошибка вычисляется по следующей формуле:

$$m_p = m_q = \sqrt{\frac{p\% \cdot q\%}{n}}$$

Для нашего примера это дает следующее:

$$m_p = m_q = \sqrt{\frac{60 \cdot 40}{100}} = \sqrt{\frac{2400}{100}} = \sqrt{24} = 4,9\%$$

Следовательно, все указанные способы вычисления дают одинаковые значения достоверности.

Следует иметь в виду, что указанный способ определения доверительных границ для долей альтернативных признаков дает правильные результаты только в тех случаях, когда **p´** и **q´,**близки к **0,5,** или находятся в границах от **0,25** до **0,75**. Если же доли приближаются к значениям **0** и **1**, то эта формула может дать неверное представление.

Поэтому, для **p´** и **q´**,имеющих величину, меньшую **0,25** и большую **0,75**, используется метод φ, предложенный Фишером. Метод φ заключается в том, что доверительные границы определяют не по методу Стьюдента, дающему распределение **t**, а с помощью величины φ, которую находят по таблицам. Величина φ имеет ошибку, независящую от величины дисперсии и величины долей **p´** и **q´**.

Величина доли **p´** связана с φ через величину синуса.

$$p' = \sin^2 \cdot \frac{\varphi}{2}$$

$$\varphi = 2\mathrm{Arcsin}\sqrt{p'}$$

Ошибка φ равна:

$$m_{\varphi} = \frac{1}{\sqrt{n}}$$

Значение величины φ находят по специальной таблице, в которой определенному значению доли **p´** соответствует значение φ.

Определение ошибки для среднего квадратичного отклонения и коэффициента изменчивости

Для различных статистических величин используются различные формулы статистических ошибок.

Статистические ошибки для **δ** и C_v выражаются следующими формулами:

Ошибка среднего квадратичного отклонения:

$$m_{\delta} = \frac{\delta}{\sqrt{2n}}$$

Ошибка коэффициента изменчивости:

$$m_{C_v} = \frac{C_v}{\sqrt{2n}}$$

Эти формулы используют для выборок с большим числом наблюдений, а критерий достоверности берут при разных уровнях вероятности:

$$t_{0,95} = 2,0$$

$$t_{0,99} = 2,6$$

$$t_{0,999} = 3,3$$

Наименьший допустимый критерий достоверности: t_δ и t_{Cv} должен быть $\geq$ **2,0 (P=0,95)**.

При необходимости сопоставления степени изменчивости двух вариационных рядов можно пользоваться не только абсолютными значениями δ_1 и δ_2 и коэффициентами вариации C_{v1} и C_{v2}, но и определять разность и ее достоверности между средними квадратичными отклонениями двух сравниваемых рядов, то есть определять

$$D = \delta_1 - \delta_2$$

Для определения достоверности разности δ_1-δ_2 вычисляют ошибку разности и критерий достоверности t_D.

Ошибку разности между средними квадратичными отклонениями двух выборок вычисляют по следующей формуле:

$$m_D = m_{\delta 1-\delta 2} = \sqrt{\frac{\delta_1^2}{2n_1} + \frac{\delta_2^2}{2n_2}},$$

где δ_1^2 и δ_2^2 - квадрат (или дисперсия) среднего квадратичного отклонения для каждой выборки;

n_1 и n_2 – число наблюдений в каждой выборке

Эту формулу используют при большом числе наблюдений **(n>30)**.

Критерий достоверности разности вычисляют обычным способом:

$$t_D = t_{\delta 1-\delta 2} = \frac{\delta_1 - \delta_2}{m_D}$$

t_D должно быть ≥2.

Пример: самок мышей опытной группы облучали лучами Рентгена, после чего сравнивали плодовитость животных этой группы с плодовитостью самок контрольной группы. Требуется определить, уменьшает или увеличивает облучение изменчивость показателя плодовитости, если получены следующие данные по группам:

Облученная группа: **n=20** голов, **δ_1=0,5** головы.

Контрольная группа: **n=15** голов, **δ_2=0,3** головы.

Определить достоверность разности **D=δ_1-δ_2=0,5-0,3=0,2** головы.

Вычисляем ошибку разности:

$$m_D = \sqrt{\frac{\delta_1^2}{2n_1} + \frac{\delta_2^2}{2n_2}} = \sqrt{\frac{0,5^2}{2 \cdot 20} + \frac{0,3^2}{2 \cdot 15}} = \sqrt{\frac{0,25}{40} + \frac{0,09}{30}} = \sqrt{0,006 + 0,003}$$
$$= \sqrt{0,009} \approx 0,09$$

Определим критерий достоверности разности:

$$t_D = \frac{\delta_1 - \delta_2}{m_D} = \frac{0,5 - 0,3}{0,09} = \frac{0,2}{0,09} = 2,22$$

Полученный критерий указывает на наличие достоверности разности при **P=0,95**.

Определение ошибки для коэффициентов асимметрии и эксцесса

Для коэффициента асимметрии **(A_s)** ошибка определяется по следующей формуле:

$$m_{A_s} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

При больших выборках это может быть упрощено:

$$\mathbf{m_{A_s}} = \sqrt{\frac{6}{n}}$$

Критерий достоверности асимметрии:

$$t_{A_s} = \frac{A_s}{m_{A_s}} \geq 3$$

Формула ошибки коэффициента эксцесса **(Ex)**:

$$\mathbf{m_{E_x}} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}$$

При большой выборке:

$$\mathbf{m_{E_x}} = \sqrt{\frac{24}{n}} = 2\sqrt{\frac{6}{n}}$$

Что дает те же результаты, если использовать удвоенную ошибку асимметрии, то есть

$$\mathbf{m_{E_x}} = 2\mathbf{m_{A_s}}$$

Критерий достоверности эксцесса выражается следующей формулой:

$$\mathbf{t_{E_x}} = \frac{\mathbf{E_x}}{\mathbf{m_{E_x}}} \geq 3$$

СТАТИСТИЧЕСКИЕ СВЯЗИ И МЕТОДЫ ВЫЧИСЛЕНИЯ ИХ ВЕЛИЧИН

Отличительной черной биологических объектов является многообразие признаков, характеризующих каждый из них. Так, животное можно характеризовать возрастом, размерами, весом, различными физиологическими показателями и т.д. Имея однородную совокупность объектов, можно изучать распределение их по любому из их признаков.

Весьма часто можно усмотреть известную связь между вариациями по различным признакам. Например, чем больше размеры животного, тем обычно больше его вес.

В простейшем случае связь между двумя переменными величинами строго однозначна. Например, вес образцов, сделанных из одного и того же материала, полностью определяется их объемом. Такого рода зависимость принято называть функциональной.

Функциональная связь – это связь между какими-либо показателями, когда при изменении одного признака или показателя на определенную величину другой признак или показатель тоже меняется на определенную

величину. Например, при повышении температуры газа на какое-то количество градусов, объем его увеличится на определенную величину.

Для биологических объектов связь обычно бывает менее «жесткой». Объекты с одинаковым значением одного признака имеют, как правило, различные значения по другим признакам. Такую связь между вариациями называют корреляцией (дословный перевод: соотношение) между признаками.

При корреляционных связях изменение одного признака у ряда особей на определенную величину будет сопровождаться изменениями другого признака на различные, то есть варьирующие значения.

Например, корреляционная связь между весом животных и их длиной выражается в том, что каждому значению длины соответствует определенное распределение веса (а не одно значение веса) и с увеличением длины увеличивается и средний вес животных.

У животных и растений все процессы и все признаки взаимно связаны и вместе с тем каждый из них, в свою очередь, связан с внешней средой, следовательно, корреляционные связи являются широко распространенными и требуют углубленного изучения.

Наиболее правильный путь изучения корреляционных связей – это определение их с использованием биологических методов, которые позволяют вскрыть природу взаимосвязи. Но, кроме того, дополнительно для выяснения величины, типа и направления связи вполне целесообразно пользоваться методом математического анализа на массе особей.

Для этого можно применять несколько статистических коэффициентами, каждый из которых позволяет выяснить различные стороны корреляционной связи.

По своим математическим особенностям корреляционные связи могут быть:

- Прямыми (положительными) и обратными отрицательными.

- Прямолинейными и криволинейными.
- Простыми и множественными.
- Между количественными признаками.
- Между качественными признаками.

Если с увеличением (или уменьшением) одного признака другой также увеличивается (или уменьшается), то такая связь называется прямой (по мере роста увеличивается вес животного).

Если с увеличением одного показателя другой будет уменьшаться, то есть изменяться в обратном направлении, то такая связь называется обратной (увеличение дозы облучения вызывает уменьшение плодовитости).

Прямолинейный – это тип связи, при котором равным друг другу изменениям одного признака соответствуют равные же изменения другого (например, с увеличением питательности кормов, увеличивается вес животных).

Криволинейный тип связи характеризуется тем, что при увеличении одного признака (или его уменьшении) другой признак сначала увеличивается, а затем уменьшается. Например, с увеличением возраста деревьев их плодоношение сначала увеличивается до определенного возраста, а затем снижается.

Простая корреляционная связь – это связь между двумя признаками, без учета имеющихся других связей (возраст – урожайность).

При множественной корреляционной связи выясняется связь между несколькими показателями (возраст, тип почвы, погодные условия - урожайность).

Корреляционные связи могут выясняться не только между количественными признаками, но и между качественными (условия произрастания - размер).

Причинность взаимосвязи между различными признаками и факторами может быть определена биологическими, биохимическими и биофизическими методами.

Коэффициенты связи, наиболее часто применяющиеся при обработке материала, следующие:

1. Коэффициент корреляции **r** для количественных и альтернативных признаков.
2. Коэффициент регрессии **R**.
3. Корреляционное отношение **η**.
4. Бисериальный показатель связи $\mathbf{r_b}$.
5. Полихорический показатель связи **ρ**.

Коэффициент корреляции r для малых и больших выборок

Коэффициент корреляции **r** позволяет определить величину и направление связи при прямолинейном ее типе или близком к прямолинейному. При криволинейной связи им пользоваться нельзя, так как он сильно преуменьшает связь, а в ряде случаев даже не может ее уловить.

Коэффициент корреляции выряжается десятичной дробью и может принимать значения от **0** до **±1**. Чем ближе значение **r** к **1**, тем больше связь между данными признаками. О тесной (сильной) связи корреляции говорят лишь тех случаях, когда коэффициент корреляции **(r)** не ниже **0,7**.

- Средняя связь **– r = 0,5-0,69**.
- Умеренная связь **– r=0,31-0,49**.
- Слабая связь **– r=0,21-0,3**.
- Очень слабая связь **r =0,2** (часто вообще не учитывается).

Формулы коэффициента корреляции могут быть выражены различно.

В общем виде формула может быть представлена как сумма произведений нормированного отклонения вариантов каждого признака от своей средней, деленной на число наблюдений:

$$\mathbf{r} = \frac{\sum[\frac{(\mathbf{V_x} - \mathbf{M_x})}{\boldsymbol{\delta}_\mathbf{x}} \cdot \frac{(\mathbf{V_y} - \mathbf{M_y})}{\boldsymbol{\delta}_\mathbf{y}}]}{\mathbf{n}},$$

или

$$\mathbf{r} = \frac{\sum \mathbf{t_x} \cdot \mathbf{t_y}}{\mathbf{n}},$$

где $\mathbf{t_x}$ и $\mathbf{t_y}$ – нормированные отклонения по признаку **x** и признаку **y**.

Использование в формуле коэффициента корреляции значений варьирующих признаков, выраженных через их нормированное отклонение, то есть в долях **δ**, позволяет вычислить связь между признаками, измеренными мерами разных именований (литры с процентами, килограммы с сантиметрами и т.п.).

Рабочие формулы коэффициента корреляции применяются с учетом того, с какой выборкой (большой или малой) и с какими значениями вариантов (однозначными, многозначными или дробными) мы имеем дело.

Так для малых выборок при многозначных показателях вариантов удобнее всего пользоваться следующими формулами:

$$\mathbf{r} = \frac{\sum \mathbf{xy} - \frac{\sum \mathbf{x} \cdot \sum \mathbf{y}}{\mathbf{n}}}{\sqrt{\mathbf{a_x} \cdot \mathbf{a_y}}},$$

где

$$\mathbf{a_x} = \sum \mathbf{x^2} - \frac{(\sum \mathbf{x})^2}{\mathbf{n}}$$

и

$$\mathbf{a_y} = \sum \mathbf{y^2} - \frac{(\sum \mathbf{y})^2}{\mathbf{n}}$$

x – варианты первого признака;

y – варианты второго признака;

n – число наблюдений в выборке.

Эта формула может быть представлена в форме, где вводятся значения средних арифметических по каждому признаку:

$$\mathbf{r} = \frac{\sum \mathbf{xy} - \mathbf{nM_x \cdot M_y}}{\sqrt{(\sum \mathbf{x^2} - \mathbf{nM_x^2})(\sum \mathbf{y^2} - \mathbf{nM_y^2})}}$$

Разберем использование этих формул для малых выборок: определить величину и направление связи между плодовитостью 10 самок ондатры материнской группы и плодовитостью их дочерей.

Плодовитость матерей x	Плодовитость дочерей y	xy	x^2	y^2
6	7	42	36	49
5	6	30	25	36
5	5	25	25	25
7	8	56	49	64
4	6	24	16	36
8	6	48	64	36
5	5	25	25	25
6	7	42	36	49
5	4	20	25	16
4	5	20	16	25
∑x=55	∑y=59	∑xy=332	$\sum x^2$=317	$\sum y^2$=361

Представим в виде вариационного ряда показатели плодовитости каждой самки, записывая парно эти данные (мать-дочь), и произведем вычисления таких показателей, как **xy**, $\mathbf{x^2}$, $\mathbf{y^2}$.

Вычислим значения $\mathbf{a_x}$ и $\mathbf{a_y}$.

$$a_x = \sum x^2 - \frac{(\sum x)^2}{n} = 317 - \frac{55^2}{10} = 317 - \frac{3025}{10} = 317 - 302,5 = 14,5$$

$$a_y = \sum y^2 - \frac{(\sum y)^2}{n} = 361 - \frac{59^2}{10} = 361 - \frac{3481}{10} = 361 - 348,1 = 12,9$$

Вычислим коэффициент корреляции:

$$r = \frac{\sum xy - \frac{\sum x \cdot \sum y}{n}}{\sqrt{a_x \cdot a_y}} = \frac{332 - \frac{55 \cdot 59}{10}}{\sqrt{14,5 \cdot 12,9}} = \frac{332 - \frac{3245}{10}}{\sqrt{187,05}} = \frac{332 - 324,5}{13,68} = \frac{7,5}{13,68}$$
$$= +0,55$$

Таким образом, связь между плодовитостью дочерей и плодовитостью их матерей значительная и положительная, то есть чем выше плодовитость матерей, тем выше плодовитость их дочерей.

Если для этого примера использовать формулу

$$r = \frac{\sum xy - nM_x \cdot M_y}{\sqrt{(\sum x^2 - nM_x^2)(\sum y^2 - nM_y^2)}},$$

то расчеты упростятся. Для этой формулы нужны значения $\sum x^2$,$\sum y^2$ и $\sum xy$, которые у нас уже вычислены.

Требуется для этой формулы найти средние арифметические плодовитости матерей M_x, дочерей M_y и их квадраты

$$M_x = \frac{\sum x}{n} = \frac{55}{10} = 5,5$$

$$M_x^2 = 5,5^2 = 30,25 \approx 30$$

$$M_y = \frac{\sum y}{n} = \frac{59}{10} = 5,9$$

$$M_y^2 = 5,9^2 = 34,81 \approx 35$$

Подставим все имеющиеся значения в формулу:

$$r = \frac{\sum xy - nM_x \cdot M_y}{\sqrt{(\sum x^2 - nM_x^2)(\sum y^2 - nM_y^2)}} = \frac{332 - 10 \cdot 5,5 \cdot 5,9}{\sqrt{(317 - 10 \cdot 30)(361 - 10 \cdot 35)}}$$

$$= \frac{332 - 324,5}{\sqrt{(317 - 300)(361 - 350)}} = \frac{7,5}{\sqrt{17 \cdot 11}} = \frac{7,5}{\sqrt{187}} = \frac{7,5}{13,68}$$

$$= +0,55$$

Таким образом, вычисление по обеим формулам дает одинаковое значение коэффициента вариации.

Для больших выборок вычисление коэффициента корреляции можно осуществлять по следующей формуле:

$$R = \frac{\sum pa_x A_y - nb_x b_y}{N \cdot \delta_x \cdot \delta_y}$$

Для обработки данных большой выборки строят корреляционную решетку, которая объединяет частоты **p** по обоим коррелирующим признакам.

Основу решетки составляют классы, получаемые из данных о варьировании каждого признака. По горизонтали записывают классы одного признака, а по вертикали – классы другого.

Пересечение столбцов и строчек классов образуют сетку решетки в виде клеток, в которые разносят данные с учетом величины обоих признаков у каждой особи, вошедшей в выборку.

Обрабатывать данные корреляционной решетки можно известным уже методом произведений или методом сумм.

Разберем на примере вычисление коэффициента вариации методом произведений: требуется вычислить коэффициент корреляции между числом эритроцитов (млн.) и содержанием гемоглобина (%) по 36 анализам крови.

x – число эритроцитов

y – содержание гемоглобина.

x	y	x	y	x	y	x	y
0,80	**22**	3,46	77	3,71	**97**	3,30	82

1,71	45	3,32	80	4,22	96	4,10	81
2,63	61	3,11	82	3,90	92	3,29	82
3,19	76	3,28	79	4,36	94	3,81	87
2,80	72	3,66	84	2,50	50	4,20	87
3,14	83	3,90	75	1,30	27	**4,47**	90
3,21	73	4,33	82	2,80	63	3,68	72
3,28	82	3,80	79	3,10	71	3,59	76
3,63	78	3,82	87	2,87	70	3,40	71

Начинаем обработку с составления классов по каждому признаку. Для этого находим минимальные и максимальные значения числа эритроцитов **(x)** и содержания гемоглобина **(y)**.

$$x_{мин.}=0,80 \qquad y_{мин.}=22$$

$$x_{мак.}=4,47 \qquad y_{мак.}=97$$

Разность по числу эритроцитов:

$$D_x=3,67$$

Разность по содержанию гемоглобина:

$$D_y= 75$$

Определяем размер класса **К** для каждого признака:

$$K_x = \frac{D_x}{8} = \frac{3,67}{8} = 0,46 \approx 0,5$$

$$K_y = \frac{D_y}{8} = \frac{75}{8} = 9,4 \approx 10$$

Строим корреляционную решетку

x	y								p_x	a_x	$p_x a_x$	$p_x a_x^2$
	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100				
0,6-1,0	. 1								1	-5	-5	25
1,1-1,5	. 1								1	-4	-4	16
1,6-2,0			. 1						1	-3	-3	9

2,1-2,5			. 1						1	-2	-2	4
2,6-3,0					… 3	. 1			4	-1	-4	4
	1 квадрат						2 квадрат					
3,1-3,5						…….. 8	….. 5		13	0	0	0
3,6-4,0						…. 4	… 3	.. 2	9	1	9	9
4,1-4,5							…. 4	.. 2	6	2	12	24
	3 квадрат						4 квадрат					
p_y	2	0	2	0	3	13	12	4	n=36	-	∑=3	∑=91
a_y	-5	-4	-3	-2	-1	0	1	2				
$p_y a_y$	-10	0	-6	0	-3	0	12	8	∑=1			
$p_y a_y^2$	50	0	18	0	3	0	12	16	∑=99			

Графы $\mathbf{p_x}$ и $\mathbf{p_y}$ заполняют суммированием частот по каждому из признаков и образуют по ним вариационные ряды. Каждый из этих вариационных рядов обрабатывают методом произведений по известному уже способу. Для этого по каждому признаку выделяют классы с условными средними **А**, которые удобнее брать исходя из центрального расположения класса и наибольшего числа частот. В данном примере классы с условной средней **А** следующие: по числу эритроцитов ($\mathbf{A_x}$) класс с градациями 3,1-3,5 млн. и по содержанию гемоглобина ($\mathbf{A_y}$) класс 71-80%.

Эти классы выделяют, отчего в решетке образуется фигура «креста»; они служат нулевыми классами, от которых остальные перечисляют по порядку в условных отклонениях $\mathbf{a_x}$ и $\mathbf{a_y}$. В сторону уменьшения признака от нулевого класса отклонения идут со знаком минус, в сторону увеличения признака – со знаком плюс.

В формуле коэффициента корреляции требуется проставить δ_x и δ_y, b_x и b_y. Эти значения определяют обычным приемом обработки вариационного ряда. Для этого заполняют графы **ра** и $\mathbf{ра}^2$ и получают их суммы:

$\sum p_x a_x = 1$ и $\sum p_x a_x^2 = 99$

$\sum p_y a_y = 3$ и $\sum p_x a_x^2 = 91$, которые входят в формулы **δ** и **b**.

В результате этого находим b_x и b_y:

$$b_x = \frac{\sum p_x a_x}{n} = \frac{1}{36} = 0,03$$

$$b_x^2 = 0,001$$

$$b_y = \frac{\sum p_y a_y}{n} = \frac{3}{36} = 0,08$$

$$b_y^2 = 0,006$$

Вычисляем δ_x и δ_y. При этом следует иметь в виду, что для формулы коэффициента корреляции средние квадратичные отклонения выражаются в относительных величинах, то есть без умножения корня на классовый промежуток **К**.

$$\delta_x = \sqrt{\frac{\sum p_x a_x^2}{n} - b_x^2} = \sqrt{\frac{99}{36} - 0,001} = \sqrt{2,749} = 1,658$$

$$\delta_y = \sqrt{\frac{\sum p_y a_y^2}{n} - b_y^2} = \sqrt{\frac{91}{36} - 0,006} = \sqrt{2,522} = 1,588$$

Для формулы коэффициента корреляции осталось неизвестным выражение $\sum p a_x a_y$. Эта сумма получается путем умножения каждого значения частот **р** по клеткам решетки на условные отклонения a_x и a_y.

При этом действие умножения осуществляется только для тех частот, которые расположены в клетках за пределами нулевых классов (за пределами «креста»), то есть в клетках 1,2,3,4 квадратов решетки.

Произведем построчное умножение $p a_x a_y$ по каждому квадрату:

1 квадрат

1-я строка: 1·-5·-5=+25

2-я строка: 1·-4·-5=+20

3-я строка: 1·-3·-3=+9

4-я строка: 1·-2·-3=+6

5-я строка: 3·-1·-1=+3

$\sum pa_x a_y$ **=63**

2 квадрат

$\sum pa_x a_y$ **=0**

3 квадрат

$\sum pa_x a_y$ **=0**

4 квадрат:

7-я строка: 3·1·1=3

2·1·2=4

8-я строка: 4·2·1=8

2·2·2=8

$\sum pa_x a_y$ **=23**

Суммарное значение по четырем квадратам: $\sum pa_x a_y$**= 63+23=+86**

Следует иметь в виду, что **1** и **4** квадраты имеют всегда положительное значение $\sum pa_x a_y$, **2** и **3 –** всегда отрицательное значение $\sum pa_x a_y$.

После вычисления $\sum pa_x a_y$ по четырем квадратам, которая в примере получилась **86**, можно произвести вычисление коэффициента корреляции:

$\sum pa_x a_y$**=+86 n=36 b_x=0,03 b_y=0,08 δ_x=1,658 δ_y =1,588**

Подставим эти числа в формулу **r**:

$$r = \frac{\sum pa_x a_y - n b_x b_y}{n \cdot \delta_x \cdot \delta_y} = \frac{86 - 36 \cdot 0,03 \cdot 0,08}{36 \cdot 1,658 \cdot 1,588} = \frac{86 - 0,086}{94,78} = \frac{85,91}{94,78} = +0,91$$

Таким образом, связь между числом эритроцитов и содержанием гемоглобина значительная и положительная (чем больше число эритроцитов, тем выше содержание гемоглобина).

По расположению частот в клетках решетки по диагонали можно судить о направлении связи, типе связи и уровне связи.

Так, если числа частот в основном группируются близко к диагонали, идущей из верхнего левого угла в правый нижний угол, то коэффициент корреляции будет иметь знак плюс.

Если частоты группируются вдоль другой диагонали решетки, то связь отрицательная и **r** будет иметь знак минус. Чем ближе группируются частоты к диагонали, тем больше значение **r.**

При распределении частот в клетках решетки беспорядочно связь будет незначительная.

Если частоты располагаются в решетке дугообразно или как бы образуют фигуру полумесяца, то связь имеет криволинейный тип и вычислять **r** нецелесообразно.

Коэффициент корреляции для альтернативных признаков r_a

При вычислении коэффициента корреляции для альтернативных признаков строят решетку, в которой два класса будут по одному признаку **(x)** и два – по другому признаку **(y).**

Формула такого коэффициента следующая**:**

$$r_a = \frac{p_1 \cdot p_4 - p_2 \cdot p_3}{\sqrt{(p_1 + p_2)(p_3 + p_4)(p_1 + p_3)(p_2 + p_4)}},$$

где p_1, p_2, p_3, p_4 – частоты в каждой клетке корреляционной решетки.

Рассмотрим это на конкретном примере: при исследовании связей между белой мастью и красными глазами у кроликов получены следующие данные, сгруппированные в таблицу:

	Красные глаза	Не красные глаза	∑
Белая шерсть	29	11	40
Окрашенная шерсть	1	59	60
∑	30	70	100

При подстановке всех значений сумм из таблицы в формулу получим:

$$r_a = \frac{p_1 \cdot p_4 - p_2 \cdot p_3}{\sqrt{(p_1 + p_2)(p_3 + p_4)(p_1 + p_3)(p_2 + p_4)}}$$

$$= \frac{29 \cdot 59 - 1 \cdot 11}{\sqrt{(29 + 1)(11 + 59)(29 + 11)(1 + 59)}} = \frac{1711 - 11}{\sqrt{30 \cdot 70 \cdot 40 \cdot 60}}$$

$$= \frac{1700}{\sqrt{5040000}} = \frac{1700}{2245} = +0,76$$

Полученный коэффициент корреляции указывает на достоверность связи.

Ошибка коэффициента корреляции

При большом числе наблюдений **(n≥100)**, и при высоком значении коэффициента корреляции ошибку вычисляют по следующей формуле:

$$m_r = \frac{1 - r^2}{\sqrt{n - 1}},$$

где **r –** коэффициент корреляции, вычисленный при **n≥100;**

n- число наблюдений в выборке.

Если объем выборки имеет меньше 100 наблюдений (**n<100**), то такое вариационное распределение коэффициента корреляции начинает

отклоняться от нормального и использование вышеприведенной формулы может дать искаженное значение ошибки.

Поэтому при малых выборках формула ошибки коэффициента корреляции видоизменяется следующим образом:

$$\mathbf{m_r = \sqrt{\frac{1 - r^2}{n - 2}}}$$

Критерий достоверности определяют по формуле:

$$\mathbf{t_r = \frac{r}{m_r}}$$

Определим ошибку коэффициента корреляции для нашего примера. Поскольку число наблюдений равно 100 ошибка коэффициента корреляции при альтернативной изменчивости определяется по формуле:

$$\boldsymbol{m_r = \frac{1 - r^2}{\sqrt{n - 1}} = \frac{1 - 0,76^2}{\sqrt{100 - 1}} = \frac{1 - 0,58}{\sqrt{99}} = \frac{0,42}{9,95} = 0,04}$$

$$\mathbf{t_r = \frac{r}{m_r} = \frac{0,76}{0,04} = 19}$$

Полученная величина критерия достоверности настолько велика, что вероятность достоверности результатов составляет более 0,999. Таким образом достоверность связи не вызывает сомнений.

Бисериальный показатель связи r_b

Бисериальный показатель связи $\mathbf{r_b}$ применяют в тех случаях, когда один признак выражен количественно, а другой имеет качественное и при том альтернативное выражение.

Формула бисериального показателя:

$$\mathbf{r_b = \frac{\frac{\sum p_+ + a}{n_+} - \frac{\sum pa}{n}}{\sqrt{\frac{\alpha}{n_+} - \frac{\alpha}{n}}}},$$

где

$$\alpha = \sum pa^2 - \frac{(\sum pa)^2}{n},$$

где **p –** частоты ряда всей выборки, распределенные по количественному признаку;

p_+ - частоты ряда по одному из альтернативных признаков **(+)**;

n – число наблюдений в выборке;

n_+ - число наблюдений в ряду одного из альтернативных признаков **(+)**;

α– величина, входящая в значение **δ** для вариационного ряда;

a – условное отклонение для ряда всей выборки по количественному признаку.

Для вычисления $\mathbf{r_b}$ строят корреляционную решетку обычным методом, в которой будут два класса для альтернативного признака (+ и -) и классы для количественного признака. Обычным способом разносят материал по классам решетки.

Пример: определить связь между плодовитостью самок мышей материнского поколения, облученных в дозе 100 рентген, и плодовитостью их дочерей, не подвергавшихся облучению.

Группы (2)	Плодовитость (голов)								p_2
	0	1	2	3	4	5	6	7	
Облученные матери (p_+)	5	10	15	15	5	-	-	-	n_+=50
Их дочери (p_-)	-	5	10	10	10	5	5	5	n_-=50
p_1	5	15	25	25	15	5	5	5	n=100
a	-3	-2	-1	0	1	2	3	4	-
p_+a	-15	-20	-15	0	5	-	-	-	$\sum p_+a$ =-45
p_1a	-15	-30	-25	0	15	10	15	20	$\sum p_1a$ =-10

p_1a^2	45	60	25	0	15	20	45	80	$\sum p_1a^2$ =290

Берем условное отклонение **a** в классе, имеющим плодовитость **3**, и выражаем в строчку остальные классы по плодовитости $\mathbf{a_0}$ вправо со знаком плюс и влево со знаком минус. Следующая строчка образуется от умножения частот материнской группы $\mathbf{(p_+)}$ на отклонение **a,** то есть составляем ряд $\mathbf{p_+a}$. Следующие две строчки составляют обычным способом для получения ряда $\mathbf{p_1a}$ и $\mathbf{p_1a^2}$. Находим **α:**

$$\alpha = \sum pa^2 - \frac{(\sum pa)^2}{n} = 290 - \frac{(-10)^2}{100} 290 - 1 = 289$$

Подставим полученные по строчкам необходимые значения в формулу бисериального показателя связи:

$$r_b = \frac{\frac{\sum p_+ + a}{n_+} - \frac{\sum pa}{n}}{\sqrt{\frac{\alpha}{n_+} - \frac{\alpha}{n}}} = \frac{\frac{-45}{50} - (\frac{-10}{100})}{\sqrt{\frac{289}{50} - \frac{289}{100}}} = \frac{-0,9 - (-0,1)}{\sqrt{5,78 - 2,89}} = \frac{-0,80}{\sqrt{2,89}} = \frac{-0,80}{1,7} = -0,47$$

Множественный и частный коэффициент корреляции

Выше изложены способы вычисления связи между двумя признаками. Но в биологии часто представляет большой интерес выяснить связь между тремя и большим числом факторов. Например, изучая какой-то лесной массив необходимо знать зависимость его продуктивности не только от возраста, но и от состава, полноты, состава почвы, условий произрастания и т.д. Следовательно, при множественной корреляции определяют величину

связи между изменениями показателя **z** при одновременных изменениях показателя **x** и **y**.

При множественной корреляции можно вычислить свободный коэффициент корреляции, который служит мерой силы связи между признаком **z** и признаками **x** и **y,** определяющими изменения **z**.

Формула свободного коэффициента связи:

$$r_{св} = +\sqrt{\frac{r_{xz}^2 - 2r_{xy} \cdot r_{yz} + r_{yz}^2}{1 - r_{xy}^2}}$$

Для которой обычным способом определяют коэффициенты корреляции между **x** и **z, x** и **y, y** и **z**.

Свободный коэффициент $r_{св}$ - число всегда положительное и изменяется от **0** до **1**.

Если $r_{св}=0$, то признак **z** не имеет линейной связи с признаками **x** и **y**.

Если же $r_{св}=1$, то связь между **z** и **x** и **y** линейная.

Свободный коэффициент используется редко, чаще требуется определять частные коэффициенты корреляции, которые позволяют выделять влияние каждого фактора из числа нескольких действующих.

Частные коэффициенты корреляции имеют следующую формулу при 3 факторах **x, y** и **z:**

$$r_{xz(y)} = \frac{r_{xz} - r_{yx} \cdot z_{yz}}{+\sqrt{(1 - r_{yx}^2)(1 - r_{yz}^2)}}$$

Это означает, что вычленяется действие на **z** только фактора **x** при изоляции фактора **y**.

В формуле:

$$r_{yz(x)} = \frac{r_{yz} - r_{xy} \cdot z_{xz}}{+\sqrt{(1 - r_{xy}^2)(1 - r_{yz}^2)}}$$

Выявляется влияние **y** на **z** при постоянстве **x**.

По такому же принципу можно составить формулу для $r_{xy(z)}$.

Частный коэффициент корреляции изменяется от **-1** до **+1**.

РЕГРЕССИЯ

Коэффициент корреляции указывает лишь на степень (тесноту) связи в изменчивости двух переменных величин, но не позволяет судить о том, как меняется одна величина по мере изменения другой. Ответ на этот вопрос дает вычисление коэффициента регрессии.

Коэффициент регрессии – величина именованная и показывает, насколько изменяется в среднем признак **x**, если коррелирующий с ним признак **y** изменяется на определенную величину.

Формула коэффициента регрессии включает в себя коэффициент корреляции и средние квадратические отклонения по обоим признакам и выражается следующим образом:

$$\mathbf{R_{xy} = r\frac{\delta_x}{\delta_y}} \text{ и } \mathbf{R_{yx} = r\frac{\delta_y}{\delta_x}}$$

Рассмотрим это на примере где мы вычисляли связь между числом эритроцитов (млн.) и содержанием гемоглобина (%) по 36 анализам крови**.**

Связь, выраженная через коэффициент корреляции равна

r=0,91

δ_x=1,658 δ_y=1,588

Подставим эти значения в формулу коэффициента регрессии:

Регрессия числа эритроцитов от содержания гемоглобина:

$$\mathbf{R_{xy} = r\frac{\delta_x}{\delta_y} = 0,91\frac{1,658}{1,588} = 0,91 \cdot 1,04 = 0,95}$$

То есть увеличение содержания гемоглобина на 1% количество эритроцитов увеличивается на 0,95 млн.

Регрессия содержания гемоглобина от числа эритроцитов:

$$\mathbf{R_{yx} = r\frac{\delta_y}{\delta_x} = 0,91\frac{1,588}{1,658} = 0,91 \cdot 0,96 = 0,87}$$

Увеличение количества эритроцитов на 1 млн. повышает содержание гемоглобина на 0,87%

Произведение обоих регрессий дает величину квадрата коэффициента корреляции, то есть

$$\mathbf{R_{xy} \cdot R_{yx} = r^2 = 0,95 \cdot 0,87 = 0,8265 = r^2}$$

$$\sqrt{0,8265} = 0,91$$

Коэффициент регрессии **R** в графическом виде означает тангенс угла наклона прямой, изображающей связь признаков в осях координат, к оси абсцисс. Для $\mathbf{R_{xy}}$ и $\mathbf{R_{yx}}$ угол наклона может быть различный.

ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ составляет своеобразный раздел биометрии, разработанный Р.Фишером.

При изучении и анализе сложных и многообразных причинно-следственных отношений между объектами и явлениями биологу приходится учитывать целый комплекс внешних и внутренних факторов, от которых, в конечном итоге зависят уровень и ход наблюдаемых процессов, те или иные биологические свойства живых организмов, их динамика и разнообразие.

При этом важно оценивать не только значение одного из факторов, но и их взаимодействие при сопряженном влиянии на популяцию и организм.

Решение подобных задач с помощью корреляционного или регрессивного анализа, как правило, не дает удовлетворительного результата.

Гораздо более удобным и совершенным статистическим приемом, позволяющим охватить весь комплекс наблюдений и процесс в целом и обладающим рядом других существенных достоинств, является метод дисперсионного анализа.

Дисперсионный анализ строится на обработке выборки, полученной по принципу случайного отбора объектов, но при этом допускается малочисленность материала и его качественная разнородность.

При дисперсионном анализе обработке подвергаются выборочные данные, оформленные в статистический комплекс.

Статистический комплекс оформляется в виде таблицы, состоящей из граф и строчек, по клеткам которой размещают сведения со значениями варьирующего признака. В этой части он схож с корреляционной таблицей.

Основное назначение дисперсионного анализа состоит в том, что он позволяет выявить статистически влияние различных факторов на изменчивость изучаемого признака.

При этом можно определять как влияние каждого фактора в отдельности, так и суммарное их воздействие, приводящее к определенной изменчивости в величине данного показателя.

Общую изменчивость или дисперсию выражают путем суммирования квадратов отклонений каждого варианта от средней арифметической, то есть в виде:

$$\sum(\mathbf{V}-\mathbf{M})^2 = \mathbf{C_y}$$

Это выражение называется общей дисперсией признака.

Общая дисперсия $\mathbf{C}_y$ может быть разложена на составные части:

- $\mathbf{C_x}$ **–** дисперсия, возникающая под влиянием различных учтенных факторов – факторная дисперсия.
- $\mathbf{C_z}$ **–** дисперсия, возникающая под влиянием различных случайных (неучтенных) факторов – остаточная дисперсия.

Следовательно, в общей форме дисперсия, то есть разнообразие и изменчивость любого признака, может быть записана:

$$\mathbf{C_y = C_x + C_z}$$

В задачу дисперсионного анализа входят вычисления и определение величины факториальной **(C_x)** и остаточной **(Cz)** дисперсии.

В самом общем виде факториальная дисперсия может быть представлена как сумма квадратов разностей между частными средними значениями признака **Мх,** получаемыми в графиках статистического комплекса по классам действующих факторов и общей арифметической **Мобщ.**, вычисляемой для всего статистического комплекса из показателей варьирующего признака.

Это можно выразить в следующем виде:

$$\mathbf{C_x = \sum (Мчастн. - Мобщ.)^2}, \quad \text{или}$$

$$\mathbf{C_x = \sum n_x\,(Мчастн - Мобщ.)^2},$$

где $\mathbf{n_x}$ **–** число наблюдений в каждом классе фактора

Случайную, или остаточную дисперсию **(Cz)** можно вычислить через сумму квадратов разностей варьирующего признака **V** по отношению к частной средней арифметической **Мчастн.**

$$\mathbf{C_z = \sum (V - Мчастн)^2}$$

Если ведут изучение изменчивости признака под влиянием нескольких факторов (кормление – **А**, возраст – **В),** то факториальная дисперсия **Cx** может быть представлена суммой из дисперсий каждого фактора раздельно **(А** и **В)** и дисперсии совместного влияния обоих факторов **(АВ).**

Это выражается в следующей форме:

$$C_x = C_A + C_B + C_{AB}$$

Общая дисперсия будет выражена такой суммой:

$$C_y = C_A + C_B + C_{AB} + C_Z$$

Куда входят частные факториальные дисперсии и остаточная дисперсия.

Чем меньше величина факториальных дисперсий и чем больше величина остаточной дисперсии, тем меньше познана изменчивость изучаемого признака при помощи дисперсионного анализа.

С помощью дисперсионного анализа можно вычислить долю или степень влияния **Cx** и **Cz** на варьирующий признак. Для этого берут отношение между дисперсиями и обозначают эти отношения через **(η^2).**

Доля влияния всех учтенных факторов на изменчивость признака выразится формулой:

$$\eta_x^2 = \frac{Cx}{Cy}$$

А доля влияния неучтенных факторов выразится формулой

$$\eta_z^2 = \frac{Cz}{Cy}$$

Если общую изменчивость принять за **1** или за **100%,** то доли ее будут составлять влияние каждого фактора на изменчивость признака:

$$\eta_y^2 = \eta_x^2 + \eta_z^2 = 1$$

Извлечение квадратного корня из этих выражений дает величину корреляционного отношения:

$$\eta_x = \sqrt{\frac{Cx}{Cy}} \text{ и } \eta_z = \sqrt{\frac{Cz}{Cy}}$$

Таким образом, в ходе дисперсионного анализа можно получить коэффициенты связи, не проводя специальной обработки выборочного материала.

Дисперсионный анализ осуществляется в несколько этапов путем обработки таблицы статистического комплекса и составлением сводной таблицы дисперсионного анализа.

Первый этап дисперсионного анализа заключается в обработке статистического комплекса для получения общей, факториальных и остаточной дисперсии **Cy, Cx, Cz**.

Второй этап состоит в вычислении долей каждой частной дисперсии в общей дисперсии, для чего вычисляются величины η_x^2 и η_z^2.

Третий этап сводится к корректированию полученных дисперсий на число степеней свободы, вычисляемых для каждой дисперсии по определенным формулам.

Корректированные дисперсии (или девианты) обозначают через δ^2 и вычисляют по формулам:

Корректированная общая дисперсия с учетом числа степеней свободы **Vy**:

$$\delta_y^2 = \frac{Cy}{Vy}$$

Корректированная факториальная дисперсия с учетом числа степеней свободы **Vx**:

$$\delta_x^2 = \frac{Cx}{Vx}$$

Корректированная остаточная дисперсия с учетом числа степеней свободы **Vz**:

$$\delta_z^2 = \frac{Cz}{Vz}$$

Если из корректированных дисперсий извлечь квадратный корень, то получим величины средних квадратичных отклонений **δ**:

$$\delta_y = \sqrt{\frac{Cy}{Vy}} = \sqrt{\frac{\sum(V - Mобщ.)^2}{Vy}}$$

$$\delta_x = \sqrt{\frac{Cx}{Vx}} = \sqrt{\frac{\sum(Мчастн. - Мобщ.)^2}{Vx}}$$

$$\delta_z = \sqrt{\frac{Cz}{Vz}} = \sqrt{\frac{\sum(V - Мобщ.)^2}{Vz}}$$

Четвертый этап дисперсионного анализа дает суждение о том, достоверно ли значение факториальной дисперсии, то есть достоверно ли влияние данного фактора **(x** или **A, B, AB** и т.п.) на варьирующий признак.

Для этой цели используют коэффициент Фишера **(F)**, который получается в результате деления факториальных дисперсий на остаточную дисперсию:

$$F = \frac{\delta_x^2}{\delta_z^2};\ F = \frac{\delta_A^2}{\delta_z^2};\ F = \frac{\delta_B^2}{\delta_z^2};\ F = \frac{\delta_{AB}^2}{\delta_z^2}$$

Для суждения о достоверности факториальных дисперсий сравнивают вычисленное значение **Fвычисл.** с величиной **Fтабл.,** которую определяют по специальным таблицам Фишера.

Если вычисленное значение **F** окажется больше или равным табличному значению **F**, то дисперсия и влияние данного фактора считают достоверными.

Типы статистических комплексов

Рабочие формулы и техника обработки выборки при дисперсионном анализе зависят от того, большая или малая выборка подвергается обработке, а также от структуры статистического комплекса.

Статистические комплексы различают по тому, сколько факторов включено в каждом из них для изучения дисперсии.

Статистические комплексы бывают:

- Однофакторными.
- Двухфакторными.
- Трехфакторными.
- С большим числом факторов.

Статистические комплексы различают между собой еще и по соотношению частот в классах факторов, входящих в них.

Статистические комплексы, имеющие больше одного фактора, бывают:

- Равномерными.
- Пропорциональными.
- Неравномерными.

В равномерных комплексах число наблюдений по классам факторов одинаковое и их отношения равны 1:1:1 и т.д.

В таблице приведен равномерный статистический комплекс, в котором рассматривается влияние двух факторов: доза облучения (**A**) и пола облучаемого животного (**B**) на показатель плодовитости (**V**)

Выборка имела малое число наблюдений:

Структура статистического комплекса включает три класса по фактору **A**: $\mathbf{A_1}$ – облучения не проводилось; $\mathbf{A_2}$ – облучение в дозе 100 рентген; $\mathbf{A_3}$ – облучение в дозе 200 рентген

И два класса по фактору **B**: $\mathbf{B_1}$ – группа самцов; $\mathbf{B_2}$ – группа самок.

В каждый класс фактора **A** входят факторы $\mathbf{B_1}$ и B_2.

Таким образом, в комплексе имеется 6 классов, или градаций.

Равномерный статистический комплекс при малом числе наблюдений **(n=18)**
(фактор **A –** доза облучения, фактор **B –** пол животного)

Фактор А (доза облучения)	A_1 =0 рентген	A_2= 100 рентген	A_3=200 рентген

Фактор В (пол животных)	B_1- самцы	B_2- самки	B_1- самцы	B_2- самки	B_1- самцы	B_2- самки
Варьирующий признак V (плодовитость)	8	10	7	6	5	4
	9	10	6	6	3	2
	10	8	8	5	5	3
Частоты p или n_x	3	3	3	3	3	3

18 подопытных мышей распределены равномерно, в каждом классе по 3 шт.; следовательно, комплекс имеет равномерный, двухфакторный тип.

В каждом классе таблицы строчки **V** проставлены сведения о плодовитости мышей. Для группы подопытных самок это данные об их плодовитости, а для самцов проставляют плодовитость слученных с ними самок.

Равномерный комплекс является частным случаем пропорционального.

Рассмотрим структуру пропорционального двухфакторного комплекса.

В таблице дано распределение 21 подопытного животного по показателю плодовитости в связи с фактором **А** (облучения) и фактора **В** (пол животных).

Соотношение частот в классах **В** для каждого класса **А** одно и то же и составляет **1:2**, хотя число наблюдений по классам различное.

Пропорциональный двухфакторный комплекс при малом числе наблюдений **(n=21)**

Фактор А (доза облучения)	A_1 =0 рентген		A_2= 100 рентген		A_3=200 рентген	
Фактор В (пол животных)	B_1- самцы	B_2- самки	B_1- самцы	B_2- самки	B_1- самцы	B_2- самки
Плодовитость (V)	10	10	8	7	6	6
	12	11	9	9	8	5
		11	6	7		7
		12		8		6

				6 10		
Частоты p или n_x	2	4	3	6	2	4
Отношение частот B по классам A	1:2		1:2		1:2	

В равномерных и пропорциональных комплексах сумма частных дисперсий равна общефакториальной дисперсии, то есть

$$C_A + C_B + C_{AB} = C_x$$

Остановимся на схеме неравномерного комплекса.

Неравномерный двухфакторный комплекс при малом числе наблюдений **(n=22)**

Фактор A (доза облучения)	A_1 =0 рентген		A_2= 100 рентген		A_3=200 рентген	
Фактор B (пол животных)	B_1- самцы	B_2- самки	B_1- самцы	B_2- самки	B_1- самцы	B_2- самки
Плодовитость (V)	10	10	8	7	6	6
	12	11	8	6	8	6
		12	6	10		5
		11	9	9		7
			6	7		
Частоты p или n_x	2	4	5	5	2	4
Отношение частот B по классам A	1:2		1:1		1:2	

Распределение частот фактора **B** по классам фактора **A** неравномерное: в классе $\mathbf{A_1}$ оно равно **1:2**, в классе $\mathbf{A_2}$ **– 1:1**, а классе $\mathbf{A_3}$ **– 1:1**.

Рабочие формулы и техника обработки статистического комплекса меняется в зависимости от его типа.

Обработка однофакторного комплекса при малом числе наблюдений

Однофакторные комплексы не бывают неравномерными, так как в их структуре представлены классы только по одному фактору.

Для вычисления общей дисперсии пользуются следующими рабочими формулами:

$$C_y = \sum V^2 - \frac{(\sum V)^2}{n}, \qquad \text{или } C_y = \sum V^2 - H,$$

где **V –** величина варьирующего признака.

Для удобства выражение $\frac{(\sum V)^2}{n}$ обозначают через **H.**

Остаточную дисперсию вычисляют по следующей формуле:

$$C_z = \sum V^2 - \sum h_x,$$

где

$$\sum h_x = \frac{(\sum V_x)^2}{n_x},$$

$\sum V_x$ - получается от суммирования варьирующего признака по каждому классу изучаемого фактора;

n_x - число наблюдений по каждому классу изучаемого фактора.

Факториальную дисперсию вычисляют по формуле:

$$C_x = \sum h_x - H$$

Разберем пример обработки однофакторного комплекса. Исходные данные и техника вычисления приведены в таблице:

Обработка однофакторного комплекса при малом числе наблюдений

Фактор А (доза облучения)	Классы по фактору А			Сводные показатели
	A_1 =0 рентген	A_2= 100 рентген	A_3=200 рентген	
Варьирующий признак V (плодовитость)	8 , 9, 10, 3,10, 10, 8, 3	7, 6, 8, 3, 6, 6, 5, 3	5, 3, 5, 3, 4, 2, 3, 3	∑V=133

V^2	64, 81, 100, 9, 100, 100, 64, 9	49, 36, 64, 9, 36, 36, 25, 9	25, 9, 25, 9, 16, 4, 9, 9	$\sum V^2$= 897
n_x	8	8	8	$\sum n_x$=24
$\sum V_x$	61	44	28	$\sum V_x$=133
$(\sum V_x)^2$	3721	1936	784	-
$h_x = \frac{(\sum V_x)^2}{n_x}$	$\frac{3721}{8} = 465$	$\frac{1936}{8} = 242$	$\frac{784}{8} = 98$	$\sum h_x$=805
$M_x = \frac{\sum V_x}{n_x}$	$\frac{61}{8} = 7{,}6$	$\frac{44}{8} = 5{,}5$	$\frac{28}{8} = 3{,}5$	$M_{общ} = \frac{\sum V}{n} = \frac{133}{24} = 5{,}5$

$$\mathbf{H = \frac{(\sum V)^2}{n} = \frac{133^2}{24} = \frac{17689}{24} = 737}$$

Суммарное значение по этой строчке дает $\sum \mathbf{V^2}$, входящее в формулу общей и остаточной дисперсии.

Строчку $\mathbf{n_x}$ составляют из числа наблюдений по каждому классу.

Строчка $\sum \mathbf{V_x}$ образуется путем суммирования величин плодовитости в каждом классе.

Строчку $\mathbf{(\sum V)^2}$ получается возведением в квадрат данных по каждому классу из предыдущей строчки.

Строчку $\mathbf{h_x}$ составляют для каждого класса из соотношений данных строчек, а именно $\mathbf{(\sum V_x)^2}$ и $\mathbf{n_x}$.

Суммарное значение по этой строчке дает величину $\sum \mathbf{h_x}$, входящую в формулу факториальной дисперсии.

Последнюю строчку комплекса составляют для получения частных средних арифметических для каждого класса и общей средней арифметической.

После обработки статистического комплекса можно приступать к дисперсионному анализу.

Методом дисперсионного анализа определим достоверность и долю влияния облучения по классам $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ и установим какова связь дозы облучения с плодовитостью животных.

Для этого вычислим дисперсии **Cy**, **Cx**, **Cz**.

$$C_y = \sum V^2 - H = \sum V^2 - \frac{(\sum V)^2}{n} = 897 - 737 = 160$$

$$C_x = \sum h_x - H = 805 - 737 = 68$$

$$C_z = \sum V^2 - \sum h_x = 897 - 805 = 92$$

Для проверки правильности расчетов произведем суммирование:

$C_y = C_x + C_z$, то есть $160 = 68 + 92$

Исходя из полученных данных о дисперсиях, составим сводную таблицу дисперсионного анализа.

В сводной таблице выделяют графы для факториальной, остаточной и общей дисперсии, что обозначается заголовками **x**, **z** и **y**.

Построчно проводят дальнейшую вычислительную работу.

В первой строчке записывают уже вычисленные значения дисперсий.

Вторая строка отводится для вычисления значений η^2, то есть выявлению доли (или проценту) влияния изучаемого фактора x и неучтенных факторов **z** на изменчивость признака **V**. Для этого вычисляют отношение каждой дисперсии к общей дисперсии.

Для проверки правильности вычисления поводят суммирование.

Дисперсии C	x	z	y
	68	92	160
Степень влияния фактора x и z на C η^2	$\eta_x^2 = \frac{C_x}{C_y} = \frac{68}{160} = 0{,}425 = 42{,}5\%$	$\eta_z^2 = \frac{C_z}{C_y} = \frac{92}{160}$ =0,575 = 57,5%	$\eta_x^2 + \eta_z^2 = \eta_y^2 = 1$
Число степеней свободы v	$v_x = l_x - 1$	$v_z = n - l_x$	$v_y = n - 1$

	$= 3 - 1 = 2$	$= 24 - 3 = 21$	$= 24 - 1 = 23$
Корректированная дисперсия δ^2	$\delta_x^2 = \frac{C_x}{v_x} = \frac{68}{2} = 34$	$\delta_z^2 = \frac{C_z}{v_z} = \frac{92}{21} = 4{,}38$	-
Коэффициент достоверности F	$\frac{\delta_x^2}{\delta_z^2} = \frac{34}{4{,}38} = 7{,}8$	-	-
Табличное значение F при vz=21, vx=2	При 0,95=3,5 При 0,99=5,8 При 0,999=9,8	-	-

Далее в третьей строчке определяют число степеней свободы **v**.

Для **Cx** число степеней свободы равно числу классов ***l*** по фактору **A** минус единица.

Для остаточной дисперсии **Cz** число степеней свободы определяют по разнице между числом наблюдений **n** и числом классов ***l***.

Число степеней свободы для общей дисперсии равно числу наблюдений **n** без единицы.

После этих вычислений можно рассчитать корректированную дисперсию (или девиату) $\boldsymbol{\delta^2}$.

Для этого каждую дисперсию (факториальную и остаточную) делят на соответствующее число классов степеней свободы **v**.

Последний этап дисперсионного анализа заключается в определении достоверности факториальной дисперсии, то есть достоверно ли влияние и доля влияния фактора на изменчивость признака.

Для этого вычисляют коэффициент достоверности Фишера (**F**) путем деления факториальной дисперсии на остаточную.

Далее сравнивают вычисленное значение **F** со значением **F** табличным.

Так как наше вычисленное **F** равно **7,8,** то можно сделать вывод, что влияние облучения на плодовитость животных достоверно при уровне вероятности **0,99**.

ЛИТЕРАТУРА

1. Глотов Н.В. Биометрия.-Л., изд-во ЛГУ, 1982.- 264 с
2. Гринин А.С. Математическое моделирование в экологии.- М.: Юнити, 2003.- 269 с.
3. Зайцев Г.Н. Математическая статистика в экспериментальной ботанике.- М.: Наука, 1987.- 424 с.
4. Ивантер Э.В. Основы практической биометрии.- Петрозаводск: изд.Карелия, 1969.- 96 с.

5. Лакин Г.ф. Биометрия .-М.: Высш.шк, 1990.-320 с.
6. Меркурьева Е.К. Биометрия в животноводстве.- М.:Колос, 1964.- 311 с.
7. Митропольский А.К. Методы статистических вычислений.- М.: Наука, 1971.- 576 с.
8. Песенко Ю.В. Принципы и методы количественного анализа в фаунистических исследованиях.- М.: Наука, 1982 .- 282 с.
9. Плохинский Н.Л. Биологическая статистика.- Минск: Высш.шк, 1973.- 320 с.
10. Терентье П.В., Ростова Н.С.Практикум по биометрии.- Л., изд-во ЛГУ, 1977.- 152 с.
11. Урбах В.Ю. Статистический анализ в биологических и медицинских исследованиях.- М.: Медицина, 1975.-295 с.
12. Яблоков А.В. Изменчивость млекопитающих.- М.: наука, 1966.- 362 с.