ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕСИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

КОНСПЕКТ ЛЕКЦИЙ

по учебной дисциплине вариативной части
математического и естественно-научного цикла

БИОМЕТРИЯ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕСИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

КОНСПЕКТ ЛЕКЦИЙ

по учебной дисциплине вариативной части
математического и естественно-научного цикла

БИОМЕТРИЯ

направление подготовки - 05.03.06 «Экология и природопользование»
профиль - «Экологическая безопасность»

Составитель:
Горбатко С.В., к.т.н.

Рассмотрен
на заседании кафедры
«Прикладная экология и охрана
окружающей среды »
Протокол № 2 от 20.09.2016

Утвержден на заседании
Учебно-издательского совета ДонНТУ
Протокол № от

Донецк – 2016

УДК 57.087.1 (076.5)

Конспект лекций по дисциплине «Биометрия» (для студентов направления подготовки 05.03.06 «Экология и природопользование», профиль «Экологическая безопасность» дневной и заочной форм обучения) / Составитель: Горбатко С.В. - Донецк: ДонНТУ 2016. – 89 с.

В конспекте представлены основные понятия «Биометрии», методы расчета показателей разнообразия признаков, законы распределения признаков в выборках.

Составитель:
С.В. Горбатко, к.т.н.

Рецензент:
Дедовец И.Г., к.т.н., доцент

Содержание

ОСНОВНЫЕ ПОНЯТИЯ БИОМЕТРИИ

Биометрия – наука о статистическом анализе групповых свойств биологических и экологических объектов. Под статистическим анализом подразумевается совокупность постулатов и методов теории вероятности и математической статистики применяемой в данном случае к особенностям биологических и экологических объектов.

В практике биометрических исследований используется своя специфическая терминология. Некоторые из этих терминов и их соответствий математическим приведены в табл.1.1.

Таблица 1.1 Математические и биометрические термины

Математика	Биометрия
1. Величина	1. Дата, признак
2. Среднее значение	2. Середня величина признака
3. Сума квадратов центральных отклонений	3. Дисперсия
4. Средний квадрат	4. Варианса
5. Рассеивание, вариабельность, разброс	5. Изменчивость

Групповые свойства делятся на две категории – основные и сопряженные.

К основным групповым свойствам относятся средний уровень признака, который характерен для всей группы в целом.

Сопряженные групповые свойства – такие групповые свойства, которые появляются вследствие развития основных свойств.

Разнообразие признака – неизбежная неодинаковость, большее или меньшее различие особей в группе по изучаемому признаку.

1. СРЕДНИЕ ВЕЛИЧИНЫ И ИХ ВИДЫ

Основным показателем групповых свойств в биометрии является средняя величина, которая широко используется в науке и практике. При изучении растений, животных, микроорганизмов и человека расчет средних показателей составляет основу обработки первичного материала.

Средние размеры особей численность и их масса служат для характеристики видов, разновидностей, сортов, пород и других биологических групп. Средние показатели физиологических процессов характеризуют интенсивность различных сторон внутреннего обмена организмов или силу действия биологических агентов и медицинских препаратов.

В производстве средние показатели стали обычными характеристиками оценки работы отдельных специалистов, хозяйств, областей.

Средняя величина какого-нибудь признака определяется для того, чтобы получить характеристику этого признака для всей изучаемой группы в целом.

Средняя величина признака определяется различными способами в зависимости от объекта наблюдения и поставленных целей. Поэтому имеется не один, а несколько видов средних, что приводит к определению типа средней: арифметической; геометрической; гармонической; квадратической и т.д., а также к общепринятому разграничению между аналитическими и позиционными средними.

К аналитическим средним относятся все средние величины, которые выражаются с помощью формул.

Позиционные – все средние, такие как медиана, разделительное значение, мода, высшее и низшее значение, которые определяются по отношению к общему распределению величин.

Для экологов и биологов наибольшее значение имеют четыре аналитических средних: средняя арифметическая, средняя геометрическая, средняя квадратическая и средняя гармоническая. Кроме того, для характеристики биологических совокупностей употребляются неаналитические позиционные средние: мода, медиана, разделительное значение, лимиты, квартили, децили и т.д.

1.1. Классификация средних величин

В общем случае средняя величина может быть отнесена к следующим категориям:
подвижная или устойчивая;
базовая или экспоненциальная или базово-экспоненциальная;
одноплоскостная или двуплоскостная;
однозначная или многозначная;
с полной или не полной областью применения.

Подвижной называется средняя, которая зависит от величины всех членов выборки, так что с изменением любого из них меняется и величина средней.

В противоположность подвижной – устойчивой средней является полусумма крайних членов. Если, например, среднеарифметическая может меняться, то полусумма крайних членов остается неизменной.

Аналитические средние делятся на базовые, экспоненциальные и базовоэкспоненциальные в зависимости от того фигурируют ли их члены выборки в математических формулах, которые их выражают, в качестве основания и/или показателя степени. Так, например, базовыми средними являются арифметическая и геометрическая.

Базовые средние могут быть одноплоскостными или двуплоскостными, в зависимости от того, входят ли члены совокупности только в числитель или знаменатель или одновременно и в числитель, и в знаменатель.

Средняя называется однозначной, если, каковы бы ни были члены выборки по их количеству и величине (положительные или отрицательные), средняя рассматриваемого типа имеет лишь одно значение.

Многозначные средние имеют несколько значений, например среднегеометрическая.

Если средняя может быть определена, какова бы ни была совокупность рассматриваемых признаков, то считается, что область применения средней полная, иначе говорят о не полной области применения.

1.2. Общая формула средних величин

Четыре основные вида средних величин можно выразить единой формулой:

$$Ср_в = \sqrt[m]{\frac{\sum_{i=1}^{n} P_i \cdot V_i^m}{\sum_{1=1}^{n} P_i}}$$

В частном случае, когда n $P_1 = P_2 = \ldots = P_n$

$$Ср = \sqrt[m]{\frac{\sum_{i=1}^{n} V_i^m}{n}}$$

где Ср, $Ср_В$ – средняя величина простая и взвешенная; V – дата (признак), отдельное значение изучаемого признака у каждого объекта исследования; m – показатель, определяющий вид средней; n – число усредняемых дат (признаков); Pi – математический вес (значимость) признака в выборке.

Придавая показателю m разные значения, например: 1, 2, -1, 0, можно получить формулы для отдельных видов средних.

При m=1 получаем формулу средней арифметической:

$$M = \frac{\sum_{i=1}^{n} V_i}{n}$$

При m=2 получаем формулу средней квадратической:

$$S = \sqrt{\frac{\sum_{i=1}^{n} V^2}{n}}$$

При m=-1 получаем формулу средней гармонической:

$$H = \frac{n}{\sum_{i=1}^{n} V_i^{-1}}$$

При m=0, после специальных преобразований, получаем формулу средней геометрической:

$$G = \sqrt[n]{\prod_{i=1}^{n} V_i}$$

Если в общую формулу средней подставить m = -∞ и m = +∞, то после преобразований получим два крайних значения в группе: min – наименьшее значение и max – наибольшее значение.

1.3. Общие свойства средних величин

Для правильного применения средних величин необходимо знать следующие свойства этих показателей: срединное расположение, абстрактность и единство суммарного действия.

По своему численному значению все средние величины занимают промежуточные положения между минимальным и максимальным значениями признака.

При этом наименьшую величину имеет средняя гармоническая, а наибольшую – средняя квадратическая. Следующая схема показывает положение каждой средней по отношению друг к другу:

$$m = \{ -\infty \; -1 \; 0 \; +1 \; +2 \; +\infty \}$$

$$Ср = \{ min < H < G < M < S < max \}$$

Учет указанных взаимоотношений между средними величинами помогает при проверке произведенных вычислений. Например, если средняя арифметическая оказалась выше максимального значения признака или если средняя геометрическая больше средней арифметической, то, очевидно, что в расчетах имеются ошибки.

Срединное расположение. Средняя признака показывает, какую величину имел бы каждый из представителей изучаемой группы, если бы все они были одинаковыми и суммарное их действие было такое же, как и от фактических не

усредненных значений этой группы. При использовании средних величин предполагается, что пока они применяются, разнородная группа заменена однородной группой, в которой все значения признака одинаковы и равны средней величине.

Например, если имеется пять значений признака: 1; 4; 5; 5; 5 со средней величиной М=4, то при использовании этой средней предполагается, что разнородная группа заменена на однородную с одинаковыми значениями: 4; 4; 4; 4; 4.

Данная особенность средних величин лежит в основе таких обычных производственных выражений как «от каждой коровы получено по 3000 л молока», «с каждого гектара собрано по 500 ц свеклы», «с каждого улья получено по 80 кг меда», «при откорме получено по 100 кг привеса на каждую голову» и т.п. Коровы дают, конечно, различные удои, на разных участках получен разный урожай и т.д., но все же для производственной характеристики хозяйства и, особенно, для плановых расчетов оказалось удобным условно принять, что все коровы дали или будут давать одинаковый удой, равный средней величине этого признака для данного стада и года («от каждой коровы»), или, что с каждого гектара получен один и тот же урожай, равный среднему урожаю с общей площади («с каждого гектара»).

Абстрактность. Заменить разнородную группу однородной можно только путем отвлечения от тех различий, которые существуют в действительности. Только абстрагируясь от имеющихся индивидуальных разнообразных значений, можно дать требуемую характеристику группы одним числом — средней величиной признака. В этом смысле всякая средняя величина есть прежде всего абстрактная величина, которая часто в действительности не существует, а иногда и не может существовать.

Например, если в университете среднее количество студентов в группе составляет 24,7 человека, то такое число вообще не может существовать в действительности эта средняя имеет вполне определенное производственное значение, например, при сравнении этого университета с другим, где аналогичная средняя равна 21,2.

Единство суммарного действия. Не всякое выравнивание различий в группе может привести к правильной средней величине. Вычисление средних величин необходимо вести таким образом, чтобы суммарное действие выровненных значений признака было бы равно суммарному действию первоначальных не усредненных значений.

Например, если четыре взрослых особи какой-нибудь промысловой птицы весили 2; 3; 3; 4 кг, то средний вес этих птиц: (2+3+3+4)/4 =3 кг.

Суммарный вес четырех усредненных значений 3 + 3 + 3 + 3=12 кг. Такой же суммарный вес имелся и в действительности: 2 + 3 + 3 + 4=12 кг. В данном случае выбор в качестве средней – средней арифметической сделан правильно, но так бывает не всегда.

Например, требуется рассчитать среднегодовой прирост популяции какого-нибудь вида за два года, если известно, что за первый год прирост

составил 20%, а за второй — 60% (от начала второго года). Используя способ средней арифметической, получаем:

$$M = \frac{20\% + 60\%}{2} = 40\%$$

В данном случае, применение этой средней не будет правильным, так как два усредненных значения в своем суммарном действии не дадут того же результата, какой дали два фактических не усредненных значения. Фактический общий суммарный прирост популяции за два года определяется следующим образом.

К концу первого года популяция составляет:

$$100\% + \frac{100 \cdot 20\%}{100\%} = 120\%$$

$$120\% + \frac{120\% \cdot 60\%}{100\%} = 192\%$$

а прирост от начала 1-го и к концу 2-го года составит 192% - 100% = 92%

Если же принять за средний прирост рассчитанную ранее величину 40%, то к концу 1-го года популяция составит:

$$100\% + \frac{100\% \cdot 40\%}{100\%} = 140\%$$

а к концу 2-го года:

$$140\% + \frac{140\% \cdot 40\%}{100\%} = 196\%$$

а прирост за два года составит 196% - 100% = 96%

Если же использовать среднюю геометрическую, то средний прирост определится следующим образом:

$$G = \sqrt[2]{120 \cdot 160} - 100 = 38{,}6\%$$

Численность популяции за 2 года:

$$100\% + \frac{100\% \cdot 38.6\%}{100\%} = 138.6\%$$

$$138{,}6\% + \frac{138.6\% \cdot 38.6\%}{100\%} = 192\%$$

Таким образом, суммарный результат будет равен фактическому: 192%-100%=92%.

1.4. Аналитические средние величины

1.4.1. Средняя арифметическая

Самым распространенным показателем среднего качества является средняя арифметическая. Вычисляется она, как указывалось ранее, по формуле:

$$M = \frac{\sum_{i=1}^{n} V_i}{n}$$

где M — средняя арифметическая, V — дата, отдельное значение изучаемого признака, n — число использованных значений признака. В развернутом виде эта формула имеет следующий вид:

$$M = \frac{V_1 + V_2 + V_3 + \ldots + V_n}{n}$$

Часто применяется при усреднении параллельных наблюдений за величиной признака, например, как масса, длина, высота тела, содержание вещества.

Пример. Три параллельных определения содержания гемоглобина в крови у одного и того же животного проводилось тремя разными лаборантами. Один измерил = 75. Другой = 80. Третий = 70.

Среднеарифметическая $M = \frac{75 + 80 + 70}{3} = 75$

1.4.2. Взвешенная средняя арифметическая

Обычно, как уже указывалось выше, чтобы рассчитать среднюю арифметическую, складывают все значения признака и полученную сумму делят на число дат. В этом случае каждое значение входит в сумму одинаковым образом, увеличивая ее на полную свою величину. Но такое не всегда является корректным. Иногда значения признака должны входить в сумму с неодинаковой (индивидуальной) поправкой. Эта поправка, выраженная определенным множителем, называется математическим весом значения.

Средняя, рассчитанная для значений признака с неодинаковыми весами, называется взвешенной средней. Взвешенная средняя арифметическая рассчитывается по следующей формуле:

$$M_{взв} = \frac{\sum_{i=1}^{n}(V_i \cdot P_i)}{\sum_{i=1}^{n} P_i}$$

где V — значение признака, дата; P — математический вес или значимость усредняемого значения. Часто P является количеством значений с данной величиной признака.

Чтобы рассчитать взвешенную среднюю арифметическую, необходимо каждое значение признака помножить на его вес, все эти произведения сложить и полученную сумму разделить на сумму весов.

Пример. Имеются результаты двух исследований веса пчел: в одном случае получена средняя величина 660 мг, в другом – 600 мг. Требуется получить общий средний вес, причем известно, что в первом исследовании был измерен вес у 100 пчел, а во втором – у 20.

В данном случае значениями признака являются средние 660 и 600 мг; их весами – численности групп p_1=100 и p_2=20. Взвешенная средняя арифметическая рассчитывается следующим образом:

$$M_{взв} = \frac{660 \cdot 100 + 600 \cdot 20}{100 + 20} = 650 \text{ мг}$$

В случае простой среднеарифметической было бы получено:

$$M = \frac{660 + 600}{2} = 630 \text{ мг},$$

что является не корректным значением.

1.4.3. Средняя геометрическая

Чтобы получить среднюю геометрическую для группы с n датами нужно все даты перемножить и из полученного произведения извлечь корень n-ой степени:

$$G = \sqrt[n]{\prod_{i=1}^{n} V_i}$$

или через логарифмическую форму:

$$G = \exp\left(\frac{1}{n} \cdot \sum_{i=1}^{n} \ln V_i\right)$$

где G – средняя геометрическая; n – число дат; П – произведение дат Vi; ln – натуральный логарифм для каждой из дат Vi.

Пример. Имеется не усредненная группа признаков:1; 4; 5; 5; 5. Найти среднегеометрическую этих величин.

$$G = \sqrt[5]{1 \cdot 5 \cdot 5 \cdot 5 \cdot 4} = 3{,}4654$$

$$\prod = 1.\ 4\ .\ 5\ .\ 5\ .\ 5 = 500$$

$$\prod_G = (3.4654)^5 = 500$$

Если некоторые из членов совокупности положительны, а другие отрицательны, то можно получить одну - две геометрические средние, а также может не быть ни одной. Если имеются два значения, то обычно принимают в качестве среднегеометрической то из них, знак которой совпадает со знаком средней арифметической.

При подсчете среднегеометрической не должно быть нулевых значений дат.

Применяется средняя геометрическая во всех случаях, когда необходимо узнать или спланировать средние приросты за определенный период. При расчетах среднего попериодного прироста возможны два основных способа применения средней геометрической.

Первый способ применяется, когда имеются сведения о приростах за каждый период, выраженных в процентах или долях (процент, деленный на 100) от начала каждого периода. В таких случаях расчет среднего прироста ведется по формуле:

$$x = \sqrt[n]{\prod_{i=1}^{n}(1 + a_i)} - 1$$

или для дат выраженных в процентах:

$$x = \sqrt[n]{\prod_{i=1}^{n}(100 + a_i)} - 100,\ \%$$

где x – средний попериодный прирост за ряд периодов равной продолжительности, a – фактический прирост за тот или иной период, выраженный в долях или процентах, n – число периодов.

Из этой формулы следует, что для нахождения среднего прироста по первому способу нужно долю фактического прироста за каждый период

прибавить к единице, полученные величины перемножить и из произведения извлечь корень n-й степени, а затем вычесть единицу.

Пример: Поголовье кроликов в хозяйстве увеличилось за 1-й год на 5%, за 2-й – на 20%, за 3-й – на 50%, за 4-й – на 50%, считая каждый раз от истекшего года. Рассчитать среднегодовой прирост популяции за эти годы.
Определим через среднегеометрическую:

$$x = \sqrt[4]{105 \cdot 120 \cdot 150 \cdot 150} - 100 = 29{,}76\%$$

Общий прирост за истекший период лет можно определить по следующей формуле:

$$G_{\Sigma} = \left(\left(1 + \frac{x}{100}\right)^{n} - 1\right) \cdot 100$$

Например, для предыдущего примера:

$$G_{\Sigma} = \left(\left(1 + \frac{29.76}{100}\right)^{4} - 1\right) \cdot 100 = 377.4\%$$

Второй способ расчета средних приростов применяется в тех случаях, когда имеются данные об абсолютных количествах особей (признаков, объектов) на начало и конец общего большого периода и требуется рассчитать средний прирост за более мелкие периоды. В таких случаях средний прирост рассчитывается по формуле:

$$x = \sqrt[n]{\frac{A_n}{A_1} - 1} \text{ или } x = \left(\sqrt[n]{\frac{A_n}{A_1} - 1}\right) \cdot 100\%$$

где x – средний прирост за более мелкие периоды. A_n – количество особей на конец общего периода, или на конец последнего n-го мелкого периода. A_1 - количество особей на начало исследуемого общего периода, или на начало 1-го мелкого периода.

Величину признака A_n (например, количество особей) на конец общего периода, или на конец последнего n-го мелкого периода лет можно определить по следующей формуле:

$$A_n = A_1 \cdot \left(1 + \frac{x}{100}\right)^{n}$$

Пример. В агропромышленном хозяйстве на начало пятилетнего периода было 100 ульев, а к концу стало 140. Определить среднегодовой процент увеличения пасеки за прошедшие пять лет.
Применяя указанную выше формулу, получим:

$$x=\sqrt[5]{\frac{140}{100}}-1=0{,}0697 \text{ или } 6{,}97\%$$

Пример. В области запланировано за пять лет увеличить объемы переработки твердых бытовых отходов (ТБО) на 60%. Требуется распределить это задание равномерно по годам.

В данном случае не заданы абсолютные количества в начале и конце общего периода, но дан общий процент прироста за весь период – 60%, что дает возможность легко получить требуемое отношение A_n/A_1. Объем переработки ТБО должен увеличиться на 60%. Это значит, что на каждые 100 единиц, бывших в начале общего периода, должно быть 160 единиц в конце. Для выполнения такого задания среднегодовой прирост можно запланировать следующим образом:

$$x=\sqrt[5]{\frac{160}{100}}-1=0.0985$$

Оказалось, что для увеличения переработки ТБО за пятилетку на 60% достаточно обеспечить среднегодовой прирост на 9,85%, а не ~ 12%:

$$\frac{60\%}{5}=12\%$$

как это могло показаться без учета того, что средний прирост образуется по принципу средней геометрической, а не средней арифметической.

1.4.4. Взвешенная средняя геометрическая

Определяется по формуле:

$$G_{взв}=\sqrt[\sum_{i=1}^{n}P_i]{\prod_{i=1}^{n}(V_i+100)^{P_i}}-100,$$

или через логарифмическую форму:

$$G_{взв}=\exp\left(\frac{1}{\sum_{i=1}^{n}P_i}\cdot\sum_{i=1}^{n}(P_i\cdot\ln V_i)\right)-100,$$

где Vi – величина признака выраженная в %.

Пример. Выборка прироста веса различных групп телят через один месяц приведена в табл.1.2. Определить средний прирост телят, используя взвешенную среднегеометрическую:

Таблица 1.2

V, %	10	14	17	21
P, шт	3	11	9	2

$$\sum_{i=1}^{n} P_i = 3+11+9+2=25$$

$$G_{взв} = \sqrt[25]{(10+100)^3 \cdot (14+100)^{11} \cdot (17+100)^9 \cdot (21+100)^2} - 100 = 15.13 \text{ кг}$$

1.4.5. Средняя и взвешенная средняя квадратическая

Средняя квадратическая вычисляется по формуле:

$$S = \sqrt{\frac{\sum_{i=1}^{n} V^2}{n}}$$

т.е. она равна корню квадратному из суммы квадратов дат, деленной на их число.

Пример. Если имеется пять дат: 1; 4; 5; 5; 5, то средняя квадратическая будет:

$$S = \sqrt{\frac{1^2 + 4^2 + 5^2 + 5^2 + 5^2}{5}} = 4.3$$

В отличие от других средних, здесь получаются наиболее завышенные значения. Употребляется средняя квадратическая, например, при расчете средних радиусов (диаметров) окружностей с целью определения средней площади чего-либо.

Пример. Измерение диаметров колоний, полученных от посева микробов определенного вида, дали следующие результаты (в мм): 15; 20; 10; 25; 30. Определить среднюю площадь посевов колоний.

Используем среднюю квадратическую:

$$S = \sqrt{\frac{15^2 + 20^2 + 10^2 + 25^2 + 30^2}{5}} = 21.22 \text{ мм}$$

Средняя площадь будет равна:

$$\bar{C} = \frac{\pi \cdot D^2}{4} = \frac{\pi \cdot 21.22^2}{5} = 353.7 \text{ мм}^2$$

Сопоставим со среднеарифметическим значением:

$$M = \frac{15+20+10+25+30}{5} = 20 \text{ мм}$$

Используя принцип единства суммарного действия проверим какая из полученных средних величин является корректной. Для этого рассчитаем общую площадь всех колоний. Общая площадь всех колоний через среднюю арифметическую составит:

$$C_s = 5 \cdot \frac{\pi \cdot M^2}{4} = 5 \cdot \pi \frac{20^2}{4} = 1570 \text{ мм}^2$$

В действительности общая площадь будет равна:

$$C_s = \frac{\pi}{4} \cdot \sum_{i=1}^{n} D^2 = n \cdot \frac{\pi}{4} \cdot S^2 = 5 \cdot \frac{\pi}{4} \cdot 21.22^2 = 1767.4 \text{ мм}$$

Таким образом, корректной средней является средняя квадратическая.

Взвешенная средняя квадратическая определяется по формуле:

$$S = \sqrt{\frac{\sum_{i=1}^{n} p_i \cdot V_i^2}{\sum_{i=1}^{n} p_i}}$$

1.4.6. Средняя гармоническая

Средняя гармоническая рассчитывается по формуле:

$$H = \frac{n}{\sum_{i=1}^{n} V_i^{-1}}$$

Например, для пяти дат 1; 4; 5; 5; 5 ее можно определить следующим образом:

$$H = \frac{5}{\frac{1}{1}+\frac{1}{4}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}} = 2.7$$

Применяется средняя гармоническая при усреднении признаков, с помощью которых в дальнейшем находят величины (другие признаки) обратно пропорциональные (или удельные) по отношению к первоначальным признакам. Например, меняющихся скоростей, с целью последующего расчета среднего или общего времени движения или протекания процесса; среднего объема с целью последующего расчета средней плотности и т.д.

В этом случае, к датам предъявляются следующие основные требования: при применении среднегармонической не должно быть нулевых дат; если имеются и отрицательные и положительные значения, необходимо чтобы знаменатель не равнялся нулю.

Пример. Почтовые голуби одной станции к месту кормежки летят со скоростью 50 км/час, а в обратном направлении – со скоростью 40 км/час. Если кроме этих данных, ничего больше неизвестно и требуется выяснить среднюю скорость полета для обоих направлений (расстояния равны), то сделать это можно, рассчитав простую среднюю гармоническую для двух дат – 50 и 40:

$$H = \frac{2}{\frac{1}{40} + \frac{1}{50}} = 44.4 \text{ км/ч}$$

1.4.7. Взвешенная средняя гармоническая

Определяется по формуле:

$$H_{взв} = \frac{\sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} \frac{P_i}{V_i}},$$

где сумму весов делят на сумму соотношений весов с соответствующими датами.

Взвешивание скоростей в этом случае производится по числителям удельных величин признаков, например, для усредняемых скоростей – по расстояниям.

Пример. Рысак на тренировках пробегал одну за другой три дистанции, различные по состоянию дороги. Скорость на первой дистанции составляла 13 км/час, на второй — 20 км/час и на третьей — 10 км/час. Известно, что первая дистанция была в 2 раза, а вторая в 4 раза длиннее третьей. По этим данным найти среднюю скорость рысака.

$$H_{взв} = \frac{2+4+1}{\frac{2}{13} + \frac{4}{20} + \frac{1}{10}} = 15.4 \text{ км/ч}$$

1.5. Простые неаналитические (позиционные) средние

Неаналитические (позиционные) средние – это такие величины рассматриваемого признака, местоположение которых в данной выборке не может быть выражено в виде аналитической функции или в виде совокупности алгебраических членов, но, которые находятся в зависимости от положения членов упорядоченной последовательности выборки, которую можно образовать. Из этих членов и могут быть определены средние положения на основе того, как они расположены по отношению к некоторым или ко всем членам совокупности.

В практике эколого-биологических исследований наибольшее применение получили следующие позиционные средние:

медиана;
квартили;
децили;
перцентили (центили);
квантили;
разделительное значение;
мода.

1.5.1. Медиана

Медиана – такое значение признака, которое разделяет монотонно возрастающую группу на две равные части. Одна часть содержит значения признака меньше чем медиана, а другая – большие значения.

Пример: 1 2 3 4 5 6 7 8 9

Медиана = 5. При этом число членов справа и слева от медианы равны по 4.

Когда число признаков нечетное выбирается центральное значение признака, которое делит группу пополам.

Если число членов четное, то любая величина, заключенная между значениями двух центральных членов делит последовательность пополам и может быть принята за медиану. Обычно за медиану принимают среднюю арифметическую двух центральных членов.

Если мы добавим к рассмотренной последовательности еще одну дату, то получим:

1 2 3 4 5 6 7 8 9 10

$$\text{Медиана} = \frac{5+6}{2} = 5{,}5$$

В общем виде, когда число членов ряда четное, медиану X определяют из решения следующего уравнения:

$$(X-V_1)\,(X-V_2)\cdot\ldots\cdot(X-V_i)=(V_{i+1}-X)\,(V_{i+2}-X)\ldots\cdot(V_n-X)$$

если n = 2, то:

$$X=\frac{V_1+V_2}{2}$$

если n = 4, то:

$$X=\frac{V_4\cdot V_3-V_2\cdot V_1}{(V_4+V_3)-(V_1+V_2)}$$

Для многочисленных групп медиану можно рассчитать по формуле:

$$M_e = W_H+k\cdot\left(\frac{\frac{n}{2}-\sum}{f}\right),$$

где Me – медиана; Wн – начала класса, в котором находится медиана; k – величина классового промежутка; n – общее число дат в группе; $\sum$ - сумма частот классов (начиная с меньшего), предшествующих классу, в котором находится медиана; f – частота класса, в котором находится медиана.

Медиана, обладая в полной мере всеми общими свойствами средних величин, дает начало целой серии позиционных величин: квартиль, дециль, перцентиль, которые носят общее название квантиль.

1.5.2. Квартили [Qi (i = 1, 2, 3)]

Если имеем монотонный возрастающий ряд признаков:

$$V_1, V_2, V_3, \ldots, V_n$$

и два целых положительных числа p и q, сумма которых равна 4, то квартилями называются конкретные величины, удовлетворяющие условию, что число членов ряда, предшествующих им, равно $\left(\frac{p}{q}\right)$ числа членов, следующих за ними. Таким образом, будем иметь:

1-я квартиль Q_1, если $\frac{p}{q}=\frac{1}{3}$, т.е. делит выборку в отношении 1 к 3;

2-я квартиль Q_2, если $\frac{p}{q}=\frac{2}{2}$. Вторая квартиль совпадает с медианой;

3-я квартиль Q3, если $\frac{p}{q}=\frac{3}{1}$. Третья делит выборку в отношении 3 к 1.

Таким образом, все три квартили разбивают выборку на 4-е равные части.

Квартили определяются следующим образом:

Если выражение $t=\frac{i\cdot n}{4}$ (i= 1,2,3) не является целым числом, то i-я квартиль является членом x_t ряда, где t — самое меньшее из целых чисел, превышающих это выражение (округленное до целого в большую сторону).

Если же отношение $\frac{i\cdot n}{4}$ - целое число, то i-я квартиль может быть представлена каждым числом, содержащимся в интервале между x_t и x_{t+l}, и, в частности, может быть равна средней арифметической этих двух членов ряда.

Пример. Если дан ряд признаков:

t=	1	2	3	4	5
V=	2	5	8	16	24

то при i = 1 (первая квартиль) величина $t=\frac{i\cdot n}{4}=\frac{1\cdot 5}{4}=1{,}25$ не является целым числом. Самое меньшее из целых чисел, превышающих это отношение $t=1{,}25\approx 2$. Итак, первая квартиль равна второму члену ряда, т.е. $Q_1 = x_2 = 5$. Аналогичным образом найдем, что вторая квартиль равна:

$$(t=\frac{i\cdot n}{4}=\frac{2\cdot 5}{4}=2.5\approx 3),\ Q_2=x_3=8,$$

а третья: $(t=\frac{i\cdot n}{4}=\frac{3\cdot 5}{4}=3.75\approx 4),\ Q_3=x_4=16,$

Пример. Если дан ряд чисел:

t=	1	2	3	4	5	6	7	8
V=	2	5	8	16	24	28	30	32

то при i = 1, 2 и 3, мы получим для отношения $t=\frac{i\cdot n}{4}$ следующие значения:

$$\frac{1\cdot 8}{4}=2 \qquad \frac{2\cdot 8}{4}=4 \qquad \frac{3\cdot 8}{4}=6$$

Таким образом, все три значения — целые.

В этом случае первая квартиль будет найдена как средняя арифметическая второго и третьего членов ряда, вторая квартиль — как средняя

арифметическая четвертого и пятого членов и третья квартиль — как средняя арифметическая шестого и седьмого членов. Таким образом будем иметь:

$$Q_1=\frac{5+8}{2}=6.5 \qquad Q_2=\frac{16+24}{2}=20 \qquad Q_3=\frac{28+30}{2}=29$$

Следует отметить, что в этом случае первая квартиль является медианой членов, предшествующих медиане всего ряда, а третья квартиль является медианой следующих за ней членов.

Пример. Если дан ряд:

t=	1	2	3	4	5	6	7
V=	2	5	8	16	24	28	30

то для отношения $t=\frac{i\cdot n}{4}$ получим следующие значения:

$$\frac{7}{4}=1.75 \quad \frac{14}{4}=3.5 \quad \frac{21}{4}=5.25$$

Таким образом, все три значения — не целые. Наименьшие целые числа, превосходящие соответственно эти три значения, будут: 2, 4, 6. Итак, квартили будут равны 2-му, 4-му и 6-му членам ряда, т. е. $Q_1 = 5$, $Q_2 = 16$, $Q_3 = 28$.

Однако этот последний случай значительно отличается от предыдущих. В них число членов ряда, предшествовавших первой квартили, составляло ровно одну треть числа членов, следовавших за ней, чего нет в данном случае. В то же время мы не видим иного способа точно определить первую квартиль. То же самое можно сказать и о третьей квартили. Однако, хотя найденные нами значения и не являются теми величинами, которые мы искали, исходя из данного нами определения квартилей, практическая необходимость заставляет нас принимать их как таковые. Впрочем, эти расхождения теряют свое значение, когда число членов ряда велико.

Как и во втором примере, в данном случае найденная нами первая квартиль является медианой членов, предшествующих медиане всего ряда, а третья квартиль — медианой следующих за ней членов.

1.5.3. Децили [D_i (i = 1, 2, 3,...,9)]

Если дан монотонно возрастающий ряд признаков:

$$V_1, V_2, V_3, \ldots, V_n$$

и два целых положительных числа p и q, сумма которых равна 10, децилями называются такие конкретные величины, которые удовлетворяют условию, что число членов ряда, предшествующих децилям, равно $\left(\frac{p}{q}\right)$ числа членов,

следующих за ними. Другими словами 9 децилей разбивают выборку на 10 равных частей следующим образом:

1-я дециль D_l, если $\frac{p}{q} = \frac{1}{9}$;

2-я » D_2, » $\frac{p}{q} = \frac{2}{8}$;

3-я » D_3, » $\frac{p}{q} = \frac{3}{7}$;

4-я » D_4, » $\frac{p}{q} = \frac{4}{6}$;

5-я » D_5, » $\frac{p}{q} = \frac{5}{5}$;

6-я » D_6, » $\frac{p}{q} = \frac{6}{4}$;

7-я » D_7, » $\frac{p}{q} = \frac{7}{3}$;

8-я » D_8, » $\frac{p}{q} = \frac{8}{2}$;

9-я » D_9, » $\frac{p}{q} = \frac{9}{1}$;.

Все децили являются членами ряда, если n - 1 кратно 10. Иначе – являются средними между пограничными членами ряда. Определяются децили таким же методом, как и квартили.

1.5.4. Центили [C_i(i = 1, 2, ..., 99)]

Если дан монотонно возрастающий ряд признаков:

$$V_1, V_2, V_3, \ldots, V_n$$

и два целых положительных числа p и q, причем p + q = 100, то центилями называются величины, удовлетворяющие условию, что число предшествующих им членов ряда равно $\left(\frac{p}{q}\right)$ числа членов, следующих за ними.

Для первой центили C_1 имеем $\left(\frac{p}{q}\right)=\frac{1}{99}$, для второй центили $\left(\frac{p}{q}\right)=\frac{1}{98}$ и т.д.
Центили находятся таким же способом, как квартили и децили.

1.5.5. Квантили [Q_{ki} (i = 1, 2,..., k-1)]

Квантили являются обобщением квартилей, децилей и центилей. Для двух целых положительных чисел p и q, причем p + q = k, квантилями называются величины, удовлетворяющие условию, что число предшествующих членов выборки равняется $\left(\frac{p}{q}\right)$ числа последующих членов. При k = 4 мы имеем квартили, при k = 10 — децили и при k = 100 — центили. Квантили находятся таким же способом, как было рассмотрено выше.

1.5.6. Разделительное значение (R_z)

Разделительным значением выборки в виде монотонно возрастающего ряда признаков называется такое среднее значение $R_z = V_k$, которое делит его на две части, удовлетворяющие условию, что сумма одной части с R_z должна превосходить сумму другой части без R_z:

$$\sum_{i=1}^{k-1} V_i < \sum_{j=k}^{n} V_j \text{, т.е. } V_1+V_2+\ldots+V_{k-1} < V_k + V_{k+1} + \ldots + V_n$$

$$\sum_{i=1}^{k} V_i > \sum_{j=k+1}^{n} V_j \text{, т.е. } V_1+V_2+\ldots+V_k > V_{k+1} + V_{k+2} + \ldots + V_n$$

$$R_z = V_k$$

Если существует равенство:

$$\sum_{i=1}^{k} V_i = \sum_{j=k+1}^{n} V_j \text{, т.е. } V_1+V_2+\ldots+V_k = V_{k+1} + V_{k+2} + \ldots + V_n,$$

то существует бесконечное множество разделительных значений, содержащихся в интервале между V_k и V_{k+1} в этом случае за разделительное значение принимают среднюю арифметическую

$$R_z=\frac{V_k+V_{k+1}}{2}$$

Пример. Рассмотрим следующий ряд признаков: 2, 4, 5, 6, 8, 10, 12
Имеем следующие неравенства:

$$2 + 4 + 5 + 6<8+10+12; \quad 17 < 30;$$

$$2 + 4 + 5 + 6 + 8 > 10 + 12; \quad 25 > 22,$$

откуда следует, что $R_z = 8$

Пример. Рассмотрим ряд значений некоторого признака: 3, 4, 5, 7, 9, 10.

Имеем: 3 + 4 + 5 + 7 = 9 + 10.

Таким образом, за разделительное значение можно принять любую величину, содержащуюся в открытом интервале между 7 и 9. Однако, как мы условились выше, будем принимать за разделительное значение среднюю арифметическую чисел, ограничивающих этот интервал = 8.

$$R_z = \frac{7+9}{2} = 8$$

1.6. Средние неаналитические взвешенные

Средние неаналитические называются взвешенными, если они подсчитываются по группам частот. Одной из основных позиционных взвешенных является мода.

1.6.1. Мода (преобладающее значение)

Модой называется такое значение, которое повторяется в исследуемой группе больше всего раз (является доминирующим значением). Это определение не исключает того практически возможного случая, когда несколько значений имеют одну и ту же максимальную частоту. Однако в большинстве случаев одно какое-либо преобладающее значение притягивает к себе особенно большое число членов ряда.

Пример:

Класс	100-119	120-139	140-159	160-179
Частота	2	20	60	15

Мода

В этом распределении наиболее многочисленным является третий класс (140—159) с частотой 60. Этот класс называют модальным.

Точное значение моды можно получить по следующей формуле:

$$M_0 = W_н + k \cdot \left(\frac{f_M - f_{M-1}}{2 \cdot f_M - f_{M-1} - f_{M+1}} \right),$$

где M_0 — мода, $W_н$ — начало модального класса, k — величина классового промежутка, $f_{м-1}$ — частота класса, предшествующего модальному, $f_м$ — частота модального класса, $f_{м+1}$ — частота класса, следующего за модальным.

Для приведенного распределения $W_н$=140, k = 10, $f_{м-1}$=20, $f_м$=60, $f_{м+1}$=15. Следовательно, мода этого распределения равна:

$$M = 140+10\cdot\left(\frac{60-20}{2\cdot 60-20-15}\right)=144.7$$

Обычно, если классы взяты не слишком мелкие (10-12 классов на всю группу), имеется всего один модальный класс. В некоторых распределениях встречается два или три модальных класса. Иногда это может быть следствием того, что в изучаемую группу попал разнородный материал, относящийся к разным категориям (более крупной и менее крупной) по изучаемому признаку.

2. ПОКАЗАТЕЛИ РАЗНООБРАЗИЯ ПРИЗНАКА

Всякая группа состоит из особей, отличающихся друг от друга по каждому из признаков. Различия эти иногда очень велики, иногда они почти незаметны, но они всегда имеются, так как невозможно найти даже двух абсолютно одинаковых особей.

При изучении общих свойств совокупностей невозможно ограничиться одними средними величинами, требуется дополнительно привлечь и такие показатели, которые характеризовали бы степень разнообразия особей в группе. Такими показателями являются:

- лимиты (lim) – максимальное (max) и минимальное (min) значение,
- среднее квадратическое отклонение (σ);
- коэффициент вариации (CV).

Кроме того, иногда употребляется квартильное и децильное отклонение.

Общим свойством показателей разнообразия является их способность характеризовать различную степень и различные особенности разнообразия.

2.1. Лимиты

Простейшим показателем разнообразия группы являются лимиты признака, т.е. имеющиеся максимум и минимум. Иногда вместе с лимитами указывается и размах признака — разность между максимальным и минимальным значениями.

Обычно, размах приписывается к лимитам в скобках: 2—7 (5).

Пример. При изучении веса быков в двух хозяйствах (а) и (б) получены следующие данные:

а) 640, 645, 650, 655, 660 M=650

б) 600, 630, 670, 680, 700 M=650

Средние живые веса быков в обоих хозяйствах одинаковы — 650 кг, однако, как видно разнообразие быков по весу во втором хозяйстве больше, чем в первом.

В этом случае, наиболее просто показать разнообразие можно при помощи лимитов и размаха:

а) lim_1 = 640 — 660 (20);

б) lim_2 = 600 —700 (100).

Как видно, оказалось, что во втором совхозе размах веса быков в пять раз больше, чем в первом.

При проведении параллельных анализов лимиты полученных результатов и их размах служат показателями качества проведенной работы. Кроме показаний степени разнообразия, лимиты дают характеристику, как достижений, так и недостатков, имеющихся в группе по изучаемому признаку.

Пример. Предположим, что сравниваются две группы каких-либо особей по длине.

1 2 3 4 5 6 7 8 9
а) 10, 11, 12, 13, 14, 15, 16, 17, 18 M=14, lim=10-18 (8)

б) 10, 14, 14, 14, 14, 14, 14, 14, 18 M=14, lim=10-18 (8)

Средние и лимиты в обеих группах одинаковы, и в то же время степень разнообразия этих групп явно различна. В первой группе все особи различны, а во второй семь особей из девяти имеют один и тот же размер. Изменчивость первой группы явно больше, чем второй, но отметить это при помощи лимитов в данном случае невозможно.

В таких и подобных им случаях, наиболее точно охарактеризовать степень разнообразия можно при помощи особого показателя — среднего квадратического отклонения.

2.2. Среднее квадратическое отклонение

Среднее квадратическое отклонение имеет совершенно исключительное значение в математической статистике и биометрии, в частности. Этот показатель используется в качестве абсолютной меры разнообразия и, кроме того, положен в основу почти всех характеристик изменчивости, распределения, корреляции, регрессии, дисперсионного анализа.

Среднее квадратическое отклонение определяется по формуле:

$$б=\sqrt{\frac{\sum_{i=1}^{n}(V_i - M)^2}{n-1}},$$

где ν = n-1 – число степеней свободы; M – средняя арифметическая (или иная средняя).

Под корнем – сумма квадратов центральных отклонений. С помощью среднеквадратического отклонения определяется степень разнообразия особей в группе по изучаемому признаку.

2.3. Число степеней свободы

Число степеней свободы равно числу элементов свободного разнообразия. Оно равно числу всех имеющихся элементов изучения без числа ограничений разнообразия.

Рассмотрим это понятие на примере. Пусть, для исследования требуется взять три объекта с любым развитием изучаемого признака. В данном случае величина признака не имеет никаких ограничений, поэтому число степеней свободы равно: ν = 3 – 0 = 3.

Если для исследования берется три объекта, но с условием, что сумма значений изучаемого признака должна быть равна определенной величине, например, 100, то первый объект может иметь признак любой величины, например, 20 (первая степень свободы), второй объект может также иметь любое значение признака, например, 30 (вторая степень свободы), третий же объект может иметь только одно определенное значение 50 (= 100 – 20 – 30) и поэтому не имеет свободы разнообразия.

Таким образом, для трех дат при одном ограничении (условии) разнообразия имеются две степени свободы (. = 3 – 1 = 2).

Для n дат при k ограничениях имеется ν=n—k степеней свободы. Например, при вычислении средней арифметической вся сумма значений признака относится к одному элементу из числа образующих эту сумму, причем никаких ограничений величины значений признака не имеется. Поэтому число элементов свободного разнообразия, образующих среднюю арифметическую, равно числу дат.

При вычислении среднего квадратического отклонения имеется одно ограничение величины признака у изучаемых объектов. Сигма вычисляется для определенной группы, имеющей определенную среднюю арифметическую. Поэтому разнообразие элементов, образующих среднее квадратическое отклонение, ограничено этим одним условием и в данном случае число степеней свободы равно числу дат без одной.

2.4. Коэффициент вариации

Среднее квадратическое отклонение является основным показателем разнообразия дат, объединяемых в изучаемые группы. При этом сигма служит непосредственным показателем разнообразия только при соблюдении следующих условий:

- сравниваются только одинаковые признаки;
- средние сравниваемых групп не должны сильно (<5%) отличаться друг от друга.

Например, если для длины зеркального карпа в одном улове M_1=28 см и σ_1=2 см, а во втором улове M_2=27 см, и σ_2=5 см, то ясно, что во втором садке разнообразие больше и рыбы менее стандартны.

Если указанные условия не выполняются и необходимо сравнивать разнообразие разных признаков или одинаковых признаков при резком различии средних, сигма непосредственно не может быть использована для сравнения разнообразия.

Пример. Имеются данные о величине среднего квадратического отклонения следующих признаков:

живой вес при рождении 3 кг
процент жира в молоке 0,2 %
живой вес взрослых коров 48 кг
высота в холке 7,2 см
удой за лактацию 600 кг

По этим данным невозможно установить, какой из указанных признаков более разнообразен. Нельзя сравнить 600 кг удоя с 7,2 см высоты в холке или с 0,2% жира и т.д.

В этом случае для сравнения разнообразия различных признаков применяется особый показатель — коэффициент вариации CV. Этот показатель является функцией обоих основных показателей — среднего квадратического отклонения и средней арифметической, выражается отвлеченным (безразмерным) числом и поэтому очень удобен для сравнения разнообразия любых призраков. Вычисляется коэффициент вариации по следующей формуле:

$$CV = \frac{\sigma}{M} \cdot 100\%$$

Например, если σ = 30 и M = 150, то:

$$CV = \frac{30}{150} \cdot 100\% = 20\%$$

Определение коэффициента вариации для приведенного выше примера вносит достаточную ясность в вопрос о том, какой из признаков более разнообразен (см. табл.2.1).

Таблица 2.1

Признак	M	σ	CV
живой вес при рождении	30 кг	3 кг	10 %
живой вес взрослых коров	400 кг	48 кг	12 %
удой за лактацию	3000 кг	600 кг	20 %
процент жира в молоке	4,0 %	0,2 %	5 %
высота в холке	120 см	7,2 см	6 %

Оказалось, что у исследованной группы животных наиболее разнообразным, изменчивым признаком является удой за лактацию, а наименее изменчивым — жирномолочность. Высота в холке менее изменчива, чем живой вес, а живой вес при рождении немного менее изменчив, чем живой вес взрослых коров.

2.5. Нормированное отклонение

Обычно степень развития признака определяется путем его измерения и выражается определенным именованным числом: 3 кг веса, 15 см длины, 4% жира в молоке, 15 кг настрига шерсти, 700 г привеса в сутки и др. Этот основной способ характеристики признаков оказывается недостаточным, когда требуется еще и оценить полученное значение, т.е. определить, можно ли его считать значительным или, наоборот, недостаточным, или находящимся в норме, выбрать лучшее и т.д.

Предположим, что из двух коров надо выбрать одну, лучшую по удою. Первая дала за 300 дней лактации 3500 кг молока, вторая в том же хозяйстве за тот же год дала 4500 кг за 300 дней лактации.

Можно ли на основании только этих данных заключить, что вторая корова лучше первой по обильномолочности? Нет, еще нельзя. При всех прочих равных условиях (оптимальные условия кормления и содержания, примерно равные периоды сухостоя, даты отела, продолжительности лактации и т. д.) коровы меняют свой удой в зависимости от возраста и пр.

Для получения полных оценок измеренных значений признаков принят особый показатель — нормированное отклонение, который рассчитывается по формуле:

$$x=\frac{V-M}{\sigma},$$

где x — нормированное отклонение; V — дата, результат непосредственного измерения признака; M — средняя арифметическая соответствующей группы, из которой взята изучаемая особь; σ — среднее квадратическое отклонение этого признака в группе.

Таким образом, нормированное отклонение показывает, на сколько сигм отклоняется значение признака от средней для соответствующей группы.

Нормированное отклонение — величина неименованная, что представляет большое удобство при сравнении развития различных признаков. При помощи нормированного отклонения можно вести сравнительную оценку особей, принадлежащих к разным видам, разным породам, возрастам, по разным признакам. При помощи нормированного отклонения можно унифицировать шкалы бонитировки животного, растительного мира, почв и т.д.

3. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ ПРИЗНАКА В ВЫБОРКАХ

Разнообразие объектов составляющих группу (выборку) это основное свойство всякой совокупности. В малочисленных группах трудно определить какую-либо закономерность в разнообразии дат. По мере увеличения численности изучаемых групп все больше проявляются закономерности в их разнообразии, которые скрыты (незаметны) в малочисленных группах.

Если имеется многочисленная группа особей, то различные значения признака встречаются в этой группе разное число раз. Это явление называется распределением признака.

При изучении эколого-биологических объектов по различным признакам можно встретить несколько типов распределения признака в изучаемой группе. В биометрических исследованиях наибольшее значение имеют следующие:

- законы распределения:
- нормальное;
- биномиальное;
- редких событий (Пуассона).

Изобразить распределение признака можно следующими основными способами:

- вариационным рядом;
- вариационной кривой;
- гистограммой;
- кумулятой.

Вариационный ряд – это упорядоченное отражение реально существующего распределения значений признака, по отдельным особям изученной группы. Вариационный ряд представляет собой двойной ряд чисел, состоящий из обозначений классов и соответствующих им частот.

Пример. Необходимо построить вариационный ряд для 1000 дат по 11-ти классам, через 20 единиц начиная со 110.

Мода (встречается наиболее часто f_{max}= 250)

Середины классов (W)	110	130	150	170	190	210	230	250	270	290	310	Сума
Частота (f)	2	20	60	160	250	240	180	70	15	2	1	1000

Медиана

Размах = Max – Min = 310 – 110 = 200;

Вариационный ряд включает в себя весь первичный материал по измерению одного какого-либо признака у всех представителей изучаемой группы. Это позволяет привести экспериментальный материал в определенный порядок и для очень многочисленных групп определить показатели, характеризующие признак, как по среднему уровню развития, так и по деталям разнообразия. Вариационный ряд позволяет без конкретных вычислений определить величину среднего уровня признака и разнообразия.

3.1. Составление вариационного ряда

При составлении вариационного ряда все величины признака разбиваются на равные интервалы – классы. Предварительно необходимо установить:

- число классов,
- величину классов,
- границы классов,
- средины классов,
- частоты по классам.

Число классов. Весь размах значений признака от минимума до максимума разделяется обычно на 8-12 равных интервалов. При точных исследованиях число классов устанавливается по следующей формуле:

$$R = 1+ 3{,}3.\ \lg(n),$$

где: n – число дат; lg(n) – логарифм десятичный от числа дат,
например: lg(100)=2.

После вычисления по этой формуле величину R округляют в большую сторону до ближайшего целого числа.

Величина классов или величина классового промежутка равна размаху значений от минимума до максимума, деленному на неокругленное число классов R.

Обычно величина классов устанавливается по формуле:

$$k= \frac{V_{max} - V_{min}}{1 + 3.3 \cdot \lg(n)}$$

где V_{max} — максимальное значение, V_{min} — минимальное значение.

Полученное дробное число при делении округляют до ближайшего целого числа. Например, если получено 43,4, то за величину классового промежутка k нужно взять 44.

Границы классов. Конец каждого класса должен быть меньше начала следующего на величину, равную принятой точности измерения (ξ).

$$W_{в(1)} = V_{max} + 0{,}5\ k - \xi \qquad W_{н(1)} = V_{max} - 0{,}5 \cdot k,$$

для i = 2…R:

$$W_{в(i)} = V_{в(i-1)} - k \qquad W_{н(i)} = V_{н(i-1)} - k$$

Например, если измеряется длина животных с точностью до ξ =1 см и установлена величина классового промежутка 5 см, то границы классов, начиная с нижнего минимального будут такими: 100—104, 105—109, 110 — 114, 115—119 и т.д.

Средины классов устанавливаются двумя способами. Если признак может быть выражен любым числом — и целым и дробным, то для установления средины класса нужно к началу класса прибавить половину классового промежутка.

В тех случаях, когда признак выражается только целыми числами, средина классов равна полусумме начала и конца класса.

Частоты классов устанавливаются путем разноски дат по классам. Обозначаются частоты классов символом f. Каждая дата, попав в соответствующий класс, приравнивается по величине ко всем другим датам, попавшим в этот класс.

3.2. Гистограмма

Гистограмма – вариационный ряд представленный в виде диаграммы, в которой различная величина частот изображается различной высотой столбцов. На гистограмме наглядно проявляются особенности распределения.

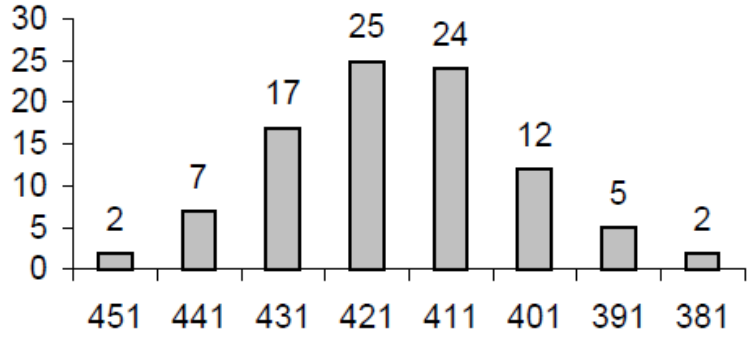


Рисунок 3.1 – Гистограмма

3.3. Вариационная кривая

Вариационная кривая (рис. 3.2) – это изображение вариационного ряда в виде кривой, ординаты которой пропорциональны частотам вариационного ряда. Вариационная кривая – это удобный и наглядный способ иллюстрации вариационного ряда в тех случаях, когда на одном графике нужно расположить или изобразить несколько распределений.

Пример. Требуется обработать результаты опытов, в которых семена помидоров подвергались облучению различными дозами рентгеновских лучей: 2х, 4х и 8-ми кратному от нормы. На контрольном (высеяны необлученные семена) и трех опытных участках, на случайно выбранных 100 кустах растений подсчитывалось число завязавшихся плодов. Результаты распределения кустов (частот) по числу завязавшихся плодов (2—4—6... 20—22) для необлученных посевов и посевов с тремя различными дозами облучения (2, 4, 8) приведены в табл.3.1. и на рисунке 3.3.

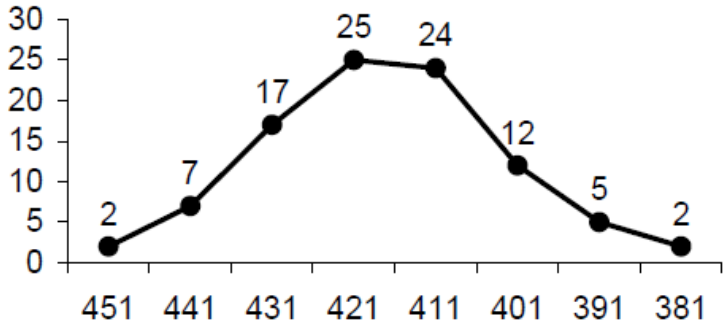


Рисунок 3.2 – Вариационная кривая

Таблица 3.1 – Результаты распределения кустов

Класс	1	2	3	4	5	6	7	8	9	10	11
W	2	4	6	8	10	12	14	16	18	20	22
0р	-	-	5	22	45	19	7	2	-	-	-
2р	-	-	4	18	42	25	8	3	-	-	-
4р	-	1	1	2	2	12	21	40	11	8	2
8р	5	33	52	8	2	-	-	-	-	-	-

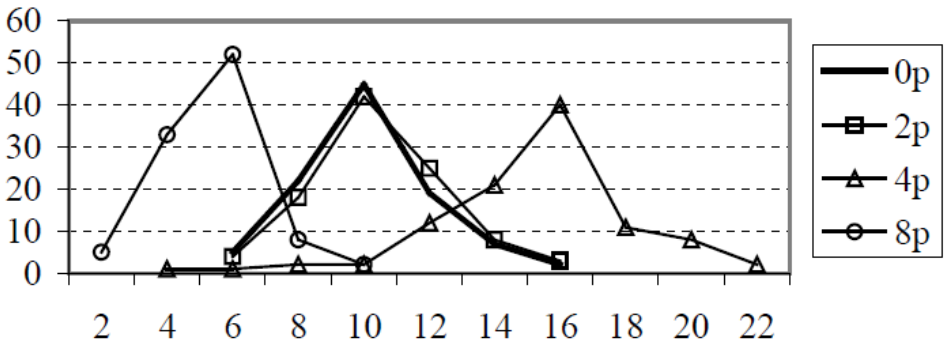


Рисунок 3.3 – Результаты распределения кустов

Сопоставление четырех вариационных кривых позволяет сделать следующий вывод: доза 2р существенно не увеличивает против контроля ни среднего числа плодов, ни разнообразия этого признака; доза 4р оказывает яв6но повышающее действие и на средний уровень, и на разнообразие (увеличивается число растений с повышенным уровнем завязавшихся плодов: 14-22); доза 8р угнетает образование плодов.

3.4. Кумулята

Кумулята – это изображение распределения признака в виде кривой, ординаты которой пропорциональны накопленным частотам вариационного ряда (рис. 3.4). Чтобы составить ряд накопленных частот, нужно к частотам наименьших классов прибавить частоту следующего класса, т.е. определить накопление суммы частот по всем классам.

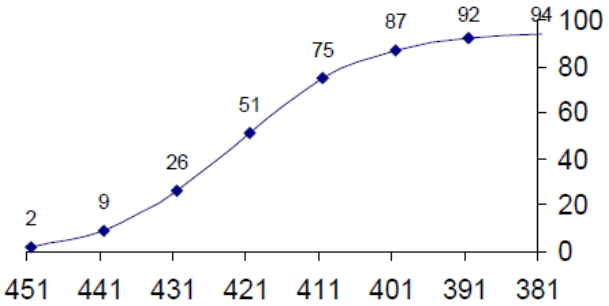


Рисунок 3.4 - Кумулята

Кумулята имеет преимущество перед вариационной кривой в случае изучения процесса накопления какого-либо признака. Один из этих методов – метод накопления (пробит) – показывает детали действия ядов и отравляющих веществ в живых организмах и в природе в целом.

3.5. Нормальное распределение

В большинстве распределений, с которыми приходится встречаться при изучении природных явлений и объектов экологу, биологу, растениеводу, зоотехнику, медицинскому работнику, проявляется определенная закономерность:

- крайние значения — наименьшее и наибольшее — появляются редко;
- чем ближе значение признака к средней арифметической, тем оно чаще встречается;
- в центре распределения имеются такие значения, которые встречаются наиболее часто и образуют в вариационном ряду модальный класс.

Данное распределение значений признака так часто встречается в самых различных областях науки и практики, что первоначально оно принималось за норму всякого массового случайного проявления признаков и в соответствии с этим получило особое название - нормальное.

В настоящее время нормальным называют распределение, которое с достаточным для практики приближением следует закону, открытому тремя учеными в разное время: Муавром в 1733 г. (Англия), Гауссом в 1809 г. (Германия) и Лапласом в 1812 г. (Франция).

Закон нормального распределения выражается следующей формулой:

$$p^* = \frac{n \cdot k}{\sigma} \cdot \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{x^2}{2}\right)$$

где: p* — теоретическая частота каждого класса распределения;
n — объем группы, число объектов исследования;
k — классовый промежуток (величина классов);

σ — среднее квадратичное отклонение: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(V_i - M)^2}{n-1}}$

$x = \frac{W - M}{\sigma}$ — нормированное отклонение средин каждого класса распределения.

Произведение в формуле 1.1: $\frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{x^2}{2}\right)$ есть f(x) — функция нормированного отклонения, которую можно рассчитать для любых значений x.

В общем случае, для определения вида распределения (нормальное, биномиальное или Пуассона) изучаемого признака выполняют сопоставление эмпирических и теоретических частот этого распределения между собой.

Нахождение ряда теоретических частот для имеющегося эмпирического распределения называется выравниванием эмпирических кривых по нормальному или другому закону, которое будет рассмотрено далее. Этот процесс имеет очень большое теоретическое и практическое значение. Выравнивание эмпирических кривых вскрывает закономерность распределения, которая обычно скрыта под случайной формой своего проявления.

Следующим важным свойством любого распределения, в том числе и нормального, является то, что можно предвидеть вероятность появления такого значения признака, которое находится в пределах заданных границ, отстоящих в обе стороны от средней на любое число сигм (средних квадратичных отклонений).

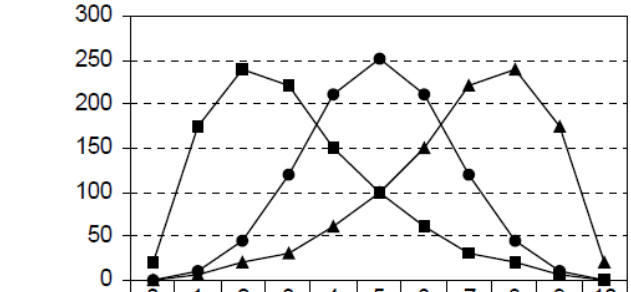
3.5.1. Асимметрия и эксцесс

Некоторые признаки у растений и животных при объединении этих объектов в группы дают распределения, значительно отличающиеся от нормального.

В тех случаях, когда какие-нибудь причины благоприятствуют появлению значений признака, отличающихся от средней величины в сторону уменьшения или в сторону увеличения, образуются асимметричные распределения. В соответствии с этим различают левую и правую асимметрии (рис. 3.5).

В тех случаях, когда какие-нибудь причины благоприятствуют преимущественному появлению и средних и крайних значений признака, образуются положительные эксцессивные распределения, имеющие вид острой пирамиды с расширенным основанием (рис. 3.6).

При отрицательном эксцессе в центре распределения имеется не вершина, а впадина, распределение становится двумодальным, вариационная кривая — двувершинной.



	0	1	2	3	4	5	6	7	8	9	10
Ряд1	1	10	45	120	210	252	210	120	45	10	1
Ряд2	20	175	240	220	150	100	60	30	20	7	1
Ряд3	1	7	20	30	60	100	150	220	240	175	20

Рисунок 3.5 – Асимметрия

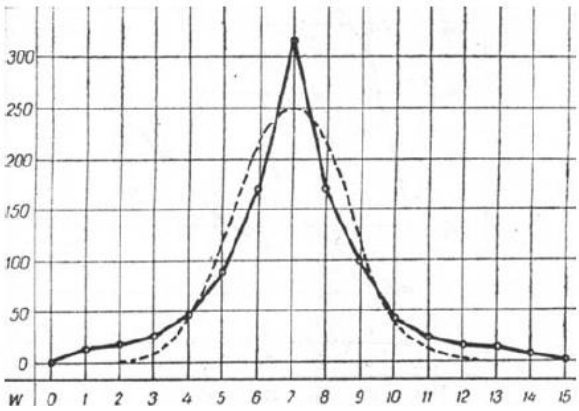


Рисунок 3.6 – Эксцесс

3.6. Достоверность различия распределений

Часто для практических и научных работ необходимо установить насколько сильно или слабо расходятся между собой эмпирические и теоретические ряды. Другими словами, возникает необходимость установить такой предел, достижение которого означает, что расхождение между эмпирическим и теоретическим (нормальным, биномиальным и т.д.)

распределением является настолько большим, что с ним необходимо считаться и данный эмпирический ряд нельзя принимать, в данном случае, за нормальный (биномиальный или другой).

В экологических (биометрических) исследованиях для этой цели применяются особые показатели – критерий χ^2 (хи-квадрат) и критерий λ (ламбда).

3.6.1. Критерий χ^2 (хи-квадрат, Пирсона)

Критерий χ^2 предложен Пирсоном и применяется во всех случаях, когда необходимо определить степень отличия фактического распределения частот от теоретического.

Определяется величина χ^2 по следующей формуле:

$$\chi^2 = \sum_{j=1}^{m} \frac{(f_j - p^*)^2}{p_j^*},$$

где: m – число классов; f — эмпирическая частота; p* — теоретическая частота.

Если крайние классы распределения имеют теоретические частоты меньше единицы, то при вычислении χ^2 их предварительно необходимо объединить в один класс вместе с ближайшим классом, имеющим частоты p*>1. Вместе с теоретическими надо объединить и соответствующие фактические частоты.

После нахождения величины χ^2 требуется определить, велика или мала она для данного распределения. Для этого пользуются таблицей предельных (стандартны) значений χ^2_{st}

Для того, чтобы пользоваться этой таблицей, необходимо предварительно установить число степеней свободы для изучаемого распределения. Если в качестве теоретического распределения берется нормальное, все детали которого определяются двумя постоянными величинами M и σ, то число степеней свободы в таких случаях равно числу классов без двух. За число классов берется то, которое получилось после объединения классов с дробными теоретическими частотами.

$$\chi^2 = \sum_{j=1}^{m} \frac{(f_j - p^*)^2}{p_j^*} \geq \chi^2_{st} \quad \{ b_1 = \text{при малой} ; b_2 – \text{при обычной}; b_3 – \text{при большой}\} \text{ ответственности исследования}$$

При выполнении этого условия расхождение считается достоверным и наблюдаемое распределение нельзя считать соответствующим принятому вначале теоретическим распределением (нормальное, биномиальное или другое).

Для каждого числа степеней свободы указаны три цифры предельных значения χ^2_{st} , соответствующие трем стандартным степеням вероятности (b_1=0,95, b_2=0,99 и b_3=0,999) того, что распределения, показавшие такие значения χ^2 или большие, различаются достоверно. Под достоверным различием понимается такое расхождение распределений, которое не может произойти в порядке обычных случайных отклонений фактических частот от теоретических.

Если фактическое значение χ^2 больше третьего значения, соответствующего b_3=0,999, то во всех случаях можно считать различие между распределениями достоверным.

3.6.2. Критерий λ (лямбда)

Критерий λ предложен советскими учеными А.Н. Колмогоровым и Н.В. Смирновым и может применяться для определения достоверности расхождения между фактическими и теоретическими распределениями, а также различий между любыми двумя распределениями частот одного и того же признака даже в том случае, когда число классов и число дат у этих распределений неодинаково. Для применения критерия λ не требуется определять число степеней свободы и не нужны таблицы для определения трех предельных значений критерия, так как для любого числа классов эти предельные значения одинаковы: 1,36; 1,63; 1,95 и соответствуют обычным трем степеням вероятности достоверного различия — b_1=0,95; b_2=0,99; b_3=0,999.

Единственным условием применения критерия λ является достаточная численность сравниваемых распределений — не менее нескольких десятков, а лучше сотен дат. Для сравнения эмпирического распределения с теоретическим при одинаковом числе классов и при одинаковой общей, численности групп критерий ламбда определяется по следующей формуле:

$$\lambda = \frac{|d|}{\sqrt{n}} = \frac{\left|\sum_{i=1}^{m} f - \sum_{i=1}^{m} p_i^*\right|_{\max}}{\sqrt{n}}$$

где d — максимальная абсолютная разность (без учета ее знака) между накопленными частотами в эмпирическом и теоретическом распределениях для одного и того же класса; n — общее число дат, образовавших эмпирическое распределение.

Для определения критерия лямбда требуется составить ряды накопленных частот (кумуляты) для обоих сравниваемых распределений $\sum f_i$ и $\sum p_i^*$, далее взять наибольшую разность (без учета ее знака) между этим величинами и полученную разность разделить на $\sqrt{n}$

Пример. При опытных посевах нового сорта пшеницы поле было разбито на 840 участков. По урожаям с каждого участка было составлено распределение с классами через 7 г/м^2. Требуется проверить, можно ли считать полученное распределение нормальным.

Для этой цели необходимо составить теоретическое нормальное распределение по полученным значениям M и σ (см. табл.3.2).

В ряду разностей (d) между накопленными частотами по обоим распределениям наибольшей величиной является 9,8. Это значение берется в качестве числителя дроби, в которой знаменатель равен корню квадратному из 840:

$$\lambda = \frac{9.8}{\sqrt{840}} = 0.34 < 1.36$$

Величина λ меньше первого предельного значения (1,36) — указывает на то, что расхождение между фактическим и теоретическим распределениями недостоверно и распределение урожая пшеницы по участкам можно считать нормальным.

Таблица 3.2 - Данные

№	W, г/м^2	f	p*	$\sum f_i$	$\sum p_i^*$	$\lvert d \rvert$
1	70	1	0.1	1	0.1	0.9
2	77	4	0.9	5	1.0	4.0
3	84	7	6.0	12	7.0	5.0
4	91	19	25.2	31	32.2	1.2
5	98	72	73.6	103	105.8	2.8
7	105	141	147.1	244	252.9	8.9
4	112	201	201.9	445	454.8	**9.8**
8	119	203	189.4	648	644.2	3.8
9	126	125	121.2	773	765.4	7.6
10	133	54	53.7	827	819.1	8.9
11	140	9	16.4	836	835.5	0.5
12	147	2	3.4	838	838.9	0.9
13	154	1	0.5	839	839.4	0.4
14	161	1	0.1	840	839.5	0.5

3.7. Биномиальное распределение

Группа особей может изучаться не только по количественным признакам, которые могут иметь различную степень своего проявления и измеряться именованными величинами — в килограммах, литрах, сантиметрах и других единицах измерения. Есть признаки, которые обычно не имеют количественных градаций (мужской пол, красная масть и др.). У каждой отдельной особи такой признак может присутствовать или отсутствовать. Такие признаки называются качественными или альтернативными.

Принципиальной разницы между количественными и качественными признаками нет. У большинства признаков, которые считаются качественными, при более тщательном изучении может быть найдена и измерена степень его

проявления, и тогда качественный признак рассматривают как количественный.

Характеристика группы по качественному признаку заключается в указании того, сколько в этой группе имеется особей с наличием данного признака и у скольких особей его нет. Для такой характеристики употребляются следующие обозначения необходимых параметров:

n — общее количество особей в группе (например, 200);

i — количество особей, имеющих изучаемый признак в группе (120);

j — количество особей, не имеющих данного признака в группе (80), (j=n–i);

$p = \frac{i}{n}$ – доля особей, имеющих признак (120/200=0,60);

$q = \frac{j}{n}$ – доля особей, не имеющих признак (80/200=0,40);

$r_{\sum} = \sum_{i=0}^{n} r_i$ – общее количество исследованных групп особей.

r_i – число групп, имеющих i-е количество особей, которые обладают рассматриваемым признаком. Очевидно следующее равенство:

$$p+q=1=\frac{\sum_{i=0}^{n} r_i}{r_{\sum}}$$

Пример. В каждом десятке из $r_{\sum}$ = 20 десятков выловленных рыб могут встретиться i = {0 (ни одной), 1, 2, 3, 4, 5, 6; 7; 8; 9 и все 10} особей, пораженных определенной болезнью. Таким образом, несколько десятков не будут иметь в своем составе пораженных рыб, несколько десятков будут иметь только по одной больной особи, несколько десятков по 2 особи и т.д.

В результате составится распределение (см. табл.3.3), в котором вариациями будут величины i — число особей, имеющих изучаемый признак в отдельных равночисленных частных группах (число больных рыб в каждом десятке), а частотами r_i – количество соответствующих равночисленных групп (число десятков).

Таблица 3.3 - Распределение

i	0	1	2	3	4	5	6	7	8	9	10
r_i	1	3	4	5	3	2	1	1	0	0	0

Полученное распределение называется биномиальным. Такое название объясняется следующими его свойствами: в распределении признак может иметь только два варианта: он есть «+» или его нет «–»; закономерности такого распределения имеют количественное выражение, связанное с коэффициентами разложения бинома Ньютона, который в применении к этому типу распределений может быть выражен следующим образом:

$$1=\frac{1}{r_{\sum}}\cdot\sum_{i=0}^{n}r_i=(p+q)^n$$

$$1=(p+q)^n=\frac{1}{r_{\sum}}\cdot\sum_{i=0}^{n}\frac{n!}{i!\cdot(n-i)!}\cdot p^i q^{n-i}$$

факториал чисел равен: 0!= 1; 1!= 1; 2!= 1. 2; 3!= 2!.3; 4!= 3!.4 и т.д.
где i – количество особей в группе, которые имеют изучаемый признак; n – общее количество особей в группе. В развернутом виде:

$$r_{\sum}\cdot(p+q)^n=\frac{1}{1}p^0\cdot q^n+\frac{n}{1}\cdot pq^{n-1}+\frac{n\cdot(n-1)}{1\cdot 2}\cdot p^2q^{n-2}+\frac{n\cdot(n-1)(n-2)}{1\cdot 2\cdot 3}p^3q^{n-3}\ldots+\frac{1}{1}\cdot p^nq^0$$

Каждый член бинома может быть представлен в виде произведения, из которых первый множитель целиком зависит от величины n:

$$f(n)=\frac{n!}{i!\cdot(n-i)!}$$

а второй – от соотношения p, q и n:

$$f(p,q)=p^i\cdot q^{n-1}$$

Подставляя в формулу бинома величины f (n) и f (p,q) , а затем решая ее относительно величины p, можно получить следующие значения:
$p^0\cdot q^n$ – нулевой член бинома (содержащий p^0 в нулевой степени), дает ожидаемую долю таких равночисленных групп, в которых из n особей ни одна не имеет изучаемого признака;
np^1q^{n-1} – первый член бинома (с p^1), дает долю групп, в которых только одна особь имеет ожидаемый признак;
$\frac{n(n-1)}{2}p^2q^{n-2}$ – второй член бинома (с p^2), дает долю групп, в которых изучаемый признак имеет по две особи;
p^n – последний член бинома, дает доля равночисленных групп, в которые все n особей имеют изучаемый признак.
Вывод:

Объемы выборок каждой группы не должны различаться между собой, т.е. должны быть равночисленными.

Для изучения необходимо брать такое количество групп, которое было бы не меньше числа особей в каждой равночисленной группе.

Пример. Среди некоторой популяции цветов семилепестковая форма цветка встречается у 10% растений. Если взять случайным образом 100 групп (букетов) по 5 растений в каждом из разных мест, то сколько можно ожидать букетов без семи лепестковой формы цветков (0) и букетов с 1, 2, 3, 4 и 5 растениями, имеющими такие цветки.

В данном случае имеется $r_{\sum}$ =100 групп, по n=5 растений в каждой, причем общая доля растений, имеющих изучаемый признак (которая была определена заранее) равна p=0,1. Определение ожидаемых частот r_i такого распределения представим в табл.3.4.

Расчеты показывают, что при общей доле p = 0,1 - растений, имеющих данный признак и r∑ = 100 групп по n = 5 растений в каждой, может встретиться 59 групп без семи лепестковой формы цветков, 33 группы, в которых из пяти растений одно будет с семилепестковыми цветками, 7 групп, имеющих из пяти два таких растения, и 1 группа, в которой из пяти особей три будут иметь изучаемый признак. Появление групп с четырьмя и пятью растениями, имеющими семилепестковую форму цветка, при данных условиях маловероятно.

Таблица 3.4 – Определение ожидаемых частот

i	f(n)	f(p,q)	$r_i=r_{\sum}\cdot f(n)\cdot f(p,q)$
0	1	$p^0q^5=1\left(\frac{9}{10}\right)^5=\frac{59049}{100000}$	59
1	5	$p^1q^4=\left(\frac{1}{10}\right)\left(\frac{9}{10}\right)^4=\frac{6561}{100000}$	33
2	10	$p^2q^3=\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^3=\frac{729}{100000}$	7
3	10	$p^3q^2=\left(\frac{1}{10}\right)^3\left(\frac{9}{10}\right)^2=\frac{81}{100000}$	1
4	5	$p^4q^1=\left(\frac{1}{10}\right)^4\left(\frac{9}{10}\right)^1=\frac{9}{100000}$	-
5	1	$p^5q^0=\left(\frac{1}{10}\right)^5 1=\frac{1}{100000}$	-

Легко подсчитать, что при $r_{\sum}$=1000 можно ожидать, что только один букет из тысячи будет иметь четыре растения с семи лепестковыми цветками, и только набрав 10000 (десять тысяч) таких букетов, можно ожидать, что среди них будет один, в котором все пять растений будут такими.

Рассмотренное ранее распределение (и следующее – распределение Пуассона) можно, с некоторой долей условности, считать частными случаями биномиального распределения. Например, для нормального p→0,5 и q→0,5

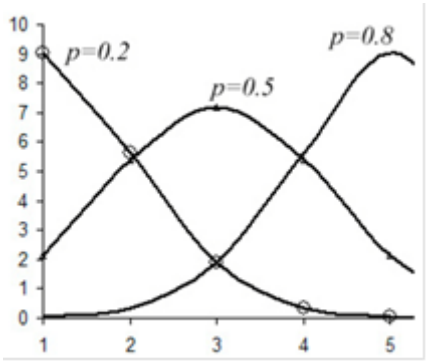


Рисунок 3.7 – Графики биномиального распределения

3.8. Распределение редких событий (Пуассона)

События, происходящие редко, один или небольшое (единичное) число раз на 1000, 10000 и большее число обычных явлений, могут быть сведены в особое распределение, в котором вариациями являются различное число редких случаев, а частотами – количество больших групп, среди которых редкое событие произошло определенное число раз.

Это распределение можно сопоставить с предыдущим биномиальным распределением, у которого p→0, а q→1 (см.рис.3).

Распределения таких редких событий обычно подчиняются определенному закону, который выражается формулой, предложенной Пуассоном:

$$p_x^* = p \cdot e^{-a} \cdot \frac{a^x}{x!}$$

где p – теоретическая частота распределения, ожидаемое число больших групп, среди которых редкое событие произошло x раз;

$p = \sum p_x$ – общее количество исследованных больших групп;

p_x – фактическая (эмпирическая) частота распределения. Число больших групп, в которых редкое событие произошло x раз;

x – число редких событий, происшедших в каждой большой группе; обычно x равно небольшому целому числу: 0, 1, 2, 3 и т. д.;

x! – произведение натуральных чисел от 1 до x (факториал). Считается, что факториал нуля равен единице: 0! = 1;

$a = \frac{\sum x \cdot p_x}{p}$ - средняя встречаемость или среднее число редких случаев на каждую большую группу. Является взвешенной средней арифметической.

Теоретическое распределение редких событий имеет одну особенность: в нем значение средней величины примерно равно квадрату сигмы (девиате). Поэтому, если в других распределениях основных величин две — M и σ, то в распределении редких событий обе основные величины сведены к одной — a (среднему числу таких событий на каждую большую группу).

Из этой особенности распределения редких событий вытекают два следствия:

1. все теоретическое распределение может быть построено на основании только одной средней – взвешенной средней арифметической;

2. при определении достоверности отличия теоретического распределения от эмпирического при помощи критерия .χ^2 число степеней свободы равно числу классов без одного.

Таблица 3.5 – Теоретически ожидаемое распределение событий

Фактическое		Теоретическое	$\chi^2=\frac{(f_i+p_i^*)^2}{p_i^*}$	$d=\left\|\sum f_i-\sum p_i^*\right\|$
x	f_x			
4	0	1	1	1
3	12	13	0.08	2
2	74	76	0.05	4
1	315	303	0.46	8
0	599	607	0.11	0
Сумма =	1000	1000	$\sum=\chi^2=1.70$ N=5-1=4 $\chi^2<\chi^2_{st}\{18.5;13.3;9.5\}$	$\lambda=$ $8/\sqrt{1000}=0.25$ $\lambda<\lambda_{st}$ {1,95; 1,63; 1,36}

где: x – число семян сорняка – повилики;

f_x – фактическое (наблюдаемое) число групп;

p^*_x - теоретическое число групп.

Пример. При проверке засоренности семян клевера оказалось, что в каждой навеске семян имелось разное количество семян повилики (опасный сорняк) — от 0 до 3. Для выяснения причин засоренности требуется определить, является ли это обычным редким явлением, вызываемым случайными обстоятельствами или нет. Для этой цели было произведено сопоставление 1000 навесок фактического распределения признака с теоретически ожидаемым распределением редких событий (см. табл.3.5).

Оказалось, что появление в семенах клевера семян повилики вполне соответствует закономерности редких явлений.

4. РЕПРЕЗЕНТАТИВНОСТЬ (ДОСТОВЕРНОСТЬ) ВЫБОРОЧНЫХ ПОКАЗАТЕЛЕЙ

Под репрезентативностью выборочных показателей понимают определение их достоверности. Обычно при любом исследовании используется два основных метода:

изучение всех особей принадлежащих данной группе или виду;

изучается только определенным образом выбранная часть.

Разница между этими двумя методами заключается в том, что в 1-м случае проводится исследование всей генеральной совокупности, во 2-м случае проводятся выборочные исследования. Остановимся на некоторых основных понятиях.

Генеральная совокупность – это весь массив особей определенной категории. Объем генеральной совокупности определяется соответствующими измерениями. Если изучается вид диких животных или растений, то генеральной совокупностью будут все особи этого вида. В этом случае объем генеральной совокупности будет очень большим и при расчетах он принимается за бесконечность (∞).

Иногда объем генеральной совокупности доступен для сплошного исследования. Если изучается небольшая совокупность необходимо определить средние величины. В этом случае генеральная совокупность может быть представлена небольшим количеством особей, но для всех исследований. Если генеральная совокупность представляет сравнительно небольшое количество особей, то она характеризуется генеральными параметрами.

Выборка – это группа объектов, которая отличается 3-мя особенностями:

представляет собой часть генеральной совокупности;

она выбирается определенным образом, но в случайном порядке;

выборка исследуется для характеристики всей генеральной совокупности.

4.1. Способы отбора объектов в выборку

Существует несколько различных способов отбора объектов в выборку. Рассмотрим каждый из них.

Случайный повторный отбор.

В этом случае объекты изучения выбираются из генеральной совокупности, но без предварительного учета развития у них изучаемых признаков, т.е. в случайном для этого признака порядке. После отбора каждый отдельный объект изучается и возвращается в свою генеральную совокупность. Таким образом, каждый объект может повторно попасть в другую выборку.

Рассмотренный способ отбора равносилен отбору из бесконечно большой генеральной совокупности, для которой разработаны основные показатели соотношений между выборочными и генеральными величинами.

Случайный бесповторный отбор.

В этом случае объекты, отобранные случайно не могут повторно попасть в данную выборку. Этот отбор является наиболее распространенным способом организации выборки. Он равносилен отбору из большой, но ограниченной генеральной совокупности, что учитывается при определении генеральных показателей по выборочным.

Механический отбор.

Производится отбор объектов из отдельных частей генеральной совокупности. Эти части предварительно намечаются механически по

квадратам опытного поля, по случайным группам животных, взятых из разных ареалов обитания популяции и т.д. Обычно намечается столько частей, сколько предполагается взять объектов для изучения, поэтому их число равно численности выборки. Механический отбор иногда осуществляется выбором для изучения особей через определенное число, например, при пропускании животных через раскол и отборе каждого десятого, сотого и т.д., или при взятии укоса через каждые 100 или 200 метров, или отборе одного объекта через каждые встретившиеся 10, 100 и т.д. экземпляров при исследовании всей популяции.

Типичный пропорциональный отбор.

Он предполагает необходимость предварительного изучения генеральной совокупности по общебиологическим или хозяйственным особенностям. На основе такого изучения вся генеральная совокупность разбивается на части, например, по типу растительных сообществ, в которых обитает вид, по рельефу местности и т.д. Из каждой такой части для изучения выбирается в случайном порядке число экземпляров, пропорциональное населенности отдельных частей.

Например, при изучении определенной породы рыб берутся уловы из разных водоемов, и при этом из каждого улова берется число экземпляров пропорциональное степени заселенности или объему водоема.

Серийный (гнездовой) отбор.

В этом случае генеральная совокупность разбивается на части (серии), некоторые из которых исследуются целиком. Этот способ применяется тогда, когда исследуемые объекты равномерно распределены либо в равном объеме, либо на равной территории.

Например, при исследовании зараженности воздуха или воды микроорганизмами для изучения берут отдельные пробы, которые подвергаются сплошному исследованию.

Поскольку часть (выборка) никогда не может полностью охарактеризовать все целое, любая характеристика генеральной совокупности на основе выборочного исследования всегда будет не точной и будет иметь некоторую большую или меньшую ошибку.

Ошибки, связанные с перенесением результатов, которые получены при изучении выборки, на всю генеральную совокупность называются ошибками репрезентативности.

Для более глубокого понимания сущности ошибок репрезентативности необходимо рассмотреть классификацию ошибок, встречающихся в научных и производственных исследованиях.

4.2. Ошибки исследований

При всяком исследовании имеется опасность допустить целый ряд ошибок самого разнообразного характера. Все эти ошибки могут быть сведены в следующие группы.

А. Общие ошибки, которые свойственны как сплошному, так и выборочному исследованиям. К ним относятся:

1. Методические ошибки. Этот класс ошибок связан со следующими действиями:

а) применение порочной методики проведения опыта (нарушение стандартных правил фиксации препаратов и химического анализа, выбор неправильного направления исследования, несоответствующего поставленным задачам, и др.);

б) не выравненность условий обитания для контрольных и опытных особей.

2. Ошибки точности. Этот класс ошибок связан со следующими действиями:

а) использование непроверенных и неправильно градуированных измерительных приборов;

б) расчеты с недостаточной точностью.

3. Случайные ошибки. С ними связаны:

а) описки, просчеты;

б) перепутывание опытных образцов.

Б. Ошибки выборочного исследования, которые свойственны только выборочным исследованиям. К ним относятся:

4. Ошибки типичности. С ними связаны:

а) отбор в выборку таких объектов, которые неправильно, односторонне отражают свойства генеральной совокупности, например, исследование только выдающихся особей или только средних или лучших, или худших;

б) отбор в выборку особей, развивавшихся в условиях, резко отличных от тех, которые характерны для всей генеральной совокупности;

в) при типическом пропорциональном отборе — отбор не из всех частей популяции и без учета объема типических частей;

г) при серийном (гнездовом) отборе — отбор не характерной серии или изучение тенденциозно выбранных особей в серии.

Все указанные категории ошибок вызываются или неправильной методикой исследования или неумелым и небрежным выполнением работы. Избежать их или свести к минимуму возможно с помощью продуманной и тщательно организованной постановке эксперимента в исследовании.

5. Ошибки репрезентативности.

Как было отмечено ранее, при выборочном исследовании существует еще один особый тип ошибок, вытекающих из самой сущности выборочного исследования и имеющих причиной то обстоятельство, что вся генеральная совокупность должна характеризоваться на основании изучения лишь ее части — выборки. Ошибок репрезентативности невозможно избежать в выборочном исследовании даже при идеальной организации исследовательской работы. Тем не менее, выборочное обследование может дать точную характеристику генеральной совокупности вследствие наличия двух благоприятных обстоятельств:

а) величину ошибок репрезентативности можно свести к минимуму определенной организацией выборочного исследования;

б) разработаны методы, позволяющие по выборочным данным определить возможную величину ошибок репрезентативности с тем, чтобы учитывать их при переходе от выборочных показателей к генеральным.

Биометрия на основе математической статистики:

- дает способы определения ошибок репрезентативности (ошибок выборочных показателей) — ошибки средней арифметической m, ошибки коэффициента корреляции m_r и др.;

- позволяет рассчитывать величину ошибок репрезентативности для выборочных показателей. Если исследуются не выборки, а генеральные совокупности, определять ошибки репрезентативности не нужно;

- определять величину ошибок репрезентативности следует только в тех случаях, когда организация исследования исключает все другие виды ошибок или когда все они сведены к минимуму.

Например, изучается вес рыб, идущих косяком, в котором обычно впереди — самки, за ними — молодь и сзади — самцы. Если в выборку попали рыбы главным образом из головной части косяка, то при определении среднего веса для всего косяка будет допущена ошибка типичности: в выборку попали особи только из одной части генеральной совокупности, отличающейся от остальных частей. В данном случае расчет ошибок репрезентативности уже не поможет, так как отбор особей в выборку произведен неправильно.

4.3. Ошибка выборочной средней арифметической

Ошибка репрезентативности средней арифметической (m) зависит от двух величин: от степени разнообразия признака (σ) в генеральной совокупности и от численности (объёма) выборки (n). Предположим, что разнообразие признака в генеральной совокупности равно нулю. Это значит, что все особи данной совокупности совершенно одинаковы. Примером может служить цвет пера у одноцветных видов птиц. В таких случаях любая выборка, даже в один экземпляр дает точную характеристику всей генеральной совокупности без какой бы то ни было ошибки репрезентативности. Чем больше разнообразие признака, тем он более изменчив, тем больше возможность попасть на такую выборку, средняя которой сильно отличается от генеральной средней. Таким образом, чем больше разнообразие, тем больше ошибка репрезентативности:

$$m \uparrow \approx \sigma \uparrow$$

Легко также понять зависимость ошибки выборочной средней от численности выборки (n). Чем больше эта численность, тем большая часть генеральной совокупности исследуется, тем с меньшей ошибкой может быть дано заключение о средней для всей генеральной совокупности.

$$m \uparrow \approx n \downarrow$$

Величина ошибки средней арифметической определяется в основном исходя из данных, полученных в выборочном исследовании. Предложено несколько формул для расчета ошибки средней — для каждого способа отбора объектов изучения.

В большинстве биологических исследований, при любом способе отбора особей в выборку можно применять единую формулу ошибки средней, формулу для случайного бесповторного отбора:

$$m = \frac{\bar{\sigma}}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

где $\bar{\sigma}$ - степень разнообразия или выборочное среднее квадратическое отклонение, полученное для изученной выборки; n – величина выборок; N – объем генеральной совокупности.

Получаемая по этой формуле величина ошибки оказывается слегка завышенной для механического, типического и серийного методов отбора. Это не представляет опасности, так как при этом получается немного более строгий подход к перенесению выборочных данных на всю генеральную совокупность.

Множитель $\sqrt{1 - \frac{n}{N}}$ при n=0 – превращает формулу ошибки средней в формулу для случайного повторного отбора, а при n=N – обращает величину ошибки в нуль.

Множитель $\sqrt{1 - \frac{n}{N}}$ оказывает влияние в тех случаях, когда в выборку попадает значительная часть генеральной совокупности, не менее 30—50%. Обычно, когда в выборке исследуется не более 5—10% особей генеральной совокупности, этот множитель настолько близок к единице, что практически не меняет значения ошибки средней и для расчета ошибки средней можно применить более простую формулу:

$$m = \frac{\bar{\sigma}}{\sqrt{n}}$$

4.4. Распределение выборочных средних

Представим, что из генеральной совокупности взято большое число отдельных выборок так, что все они исчерпали всю генеральную совокупность: $\sum n = N$. Каждая из этих выборок будет иметь свою среднюю арифметическую. Все эти средние величины не одинаковы и из них можно составить распределение, в которое в качестве дат войдут выборочные средние.

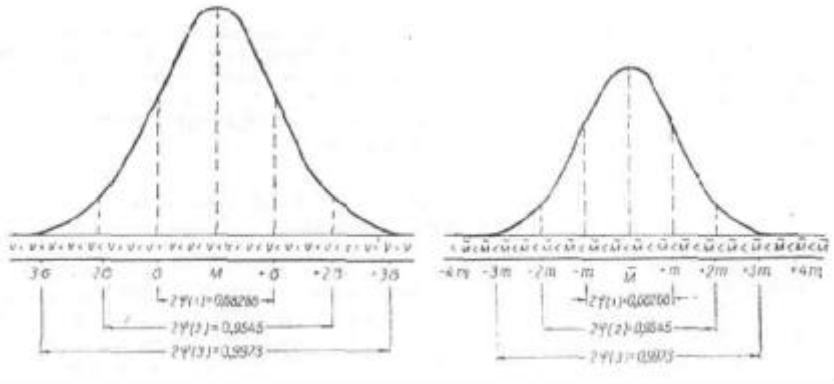


Рисунок 4.1 - Распределение выборочных средних

Как доказывается в математической статистике, средняя величина этого распределения будет равна генеральной средней, а среднее квадратическое отклонение этого распределения выборочных средних будет равно ошибке выборочной средней, приближенную величину которой рассчитывают по приведенным выше формулам.

Таким образом, рассчитывая величину ошибки выборочной средней, тем самым определяют с достаточной для практики точностью сигму (σ) ряда, составленного из выборочных средних для всех выборок из общей изучаемой генеральной совокупности.

Очень важным свойством распределения таких выборочных средних является то, что распределение имеет достаточную близость к нормальному, даже в тех случаях, когда распределение индивидуальных дат в генеральной совокупности отличается от нормального.

Это означает, что в границах от M $\pm\bar{\sigma}$ имеется 68,3% дат, которые по своему значению не отличаются от средней величины более чем на $\pm\bar{\sigma}$.Возможность ошибки данного утверждения равна 31,7% или 1 из 3.

Если за границы взять M ± $2\bar{\sigma}$, то окажется, что процент дат равен 95,4%, которые не отличающихся от средней не более чем на ± $2\bar{\sigma}$. Здесь возможность ошибки равна 4,6% или 1 из 22.

Если в качестве границ значений признака взятьM ± $3\bar{\sigma}$, то окажется, что внутри этих границ в нормальном распределении имеется 99,7% дат. Здесь возможность ошибки равна 0,3% или 1 из 370. Она считается достаточно малой и в большинстве исследований ею можно пренебречь.

На основе этих соображений и было введено в практику исследований «правило трех сигм», считающееся единым критерием границ, условно включающих в себя все распределение.

4.5. Три степени вероятности безошибочного прогноза при определении генеральных величин по выборочным

Многолетняя и обширная практика применения методов математической статистики в экологических, биологических и других исследованиях показала,

что пределы допустимых границ по «правилу трех сигм» слишком велики и исходят из требований излишней осторожности.

В большинстве случаев можно расширить границы, условно вмещающие все распределение, приняв за такие границы $M \pm 2\bar{\sigma}$.

В настоящее время применяются три степени вероятности того, что заключение о границах, вмещающих все распределение, не будет ошибочным:

$$M \pm 2\bar{\sigma}\ ;\ M \pm 2{,}5\bar{\sigma}\ ;\ M \pm 3\bar{\sigma}$$

или

$$M \pm 2m\ ;\ M \pm 2{,}5m\ ;\ M \pm 3m$$

В соответствии с этим устанавливаются три стандартных фиксированных значения, или порога вероятности безошибочного прогноза:

$$x=\frac{V-M}{y} \qquad t=\frac{\overline{M}-M}{m}$$

порог t=2 - вероятность безошибочного прогноза допускается для большинства исследовании в общей биологии, цитологии, физиологии, генетике, ботанике, зоологии, медицине (ошибка равна 100 – 95,4 = 4,6%, т.е. 1 из 22).

порог t=2,5 - (98,8%) дает возможность ошибиться 1 из 81. Такая вероятность безошибочного прогноза требуется в экономических исследованиях, связанных с рекомендациями проведения затрат денежных средств и труда, а также при обоснованиях реорганизации производства.

порог t=3 - такая вероятность требуется в особо ответственных работах: в исследованиях, проверяющих спорные теоретические выводы, в экспериментах, выясняющих вредное действие веществ и др. (ошибка равна 100 – 99,7 = 0,3%, т.е. 1 из 370).

В дальнейшем за показатели границ, условно включающих все распределение, стали приниматься округленные значения вероятностей для:
1-й степени 0,950;
2-й — 0,990;
3-й — 0,999.

В соответствии с этим были уточнены и значения t для:
1-й степени t_1=1,96 (а не 2);
2-й — t_2=2,58 (а не 2,5);
3-й — t_3=3,30 (а не 3).

Рассмотренные степени вероятности справедливы только для таких исследований, которые имеют дело с достаточно многочисленными выборками. При малочисленных выборках распределение выборочных средних, а также всех выборочных величин уже достаточно сильно отличается от нормального и следует закону распределения малых выборок,

установленному английским ученым Госсетом, писавшим под псевдонимом Student (Студент).

Распределение Стьюдента отличается от нормального тем больше, чем меньше численность выборки, причем для каждой численности малой выборки имеется свое частное распределение.

Для каждого значения численности малых выборок можно заранее рассчитать величину порога t для трех принятых степеней вероятности.

Например, при 1-й степени вероятности (b_1=0,95) и при численности выборки n=10 показатель вероятности t_1=2,3, а при численности n=3 – показатель t_1=4,3.

Значение величины порога t для любой численности выборок и для трех степеней вероятности безошибочного прогноза находят приближенно по формуле:

$$t_v = t_\infty \cdot \frac{t_\infty}{\nu + 3 - 1.5 \cdot t_\infty}$$

где t_v -показатель вероятности для выборок с числом степеней свободы – ν; t_∞ — показатель вероятности для больших выборок. В зависимости от типа ответственности исследований он равен или t_1=1,96 (b_1=0,95), или t_2=2,58 (b_2=0,99), или t_3=3,30 (b_3=0,999); ν — число степеней свободы. При определении генеральной средней по выборочной v=n - 1, при определении достоверности разности средних для некоррелированных (не взаимосвязанных) выборок $v=n_1+n_2-2$.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Часто во многих исследованиях требуется изучить несколько признаков в их взаимной связи. Если вести такое исследование по отношению к двум признакам, то можно заметить, что изменчивость одного признака находится в некотором соответствии с изменчивостью другого. В некоторых случаях такая зависимость проявляется настолько сильно, что при изменении первого признака на определенную величину всегда изменяется и второй признак на определенную величину, поэтому каждому значению первого признака всегда соответствует совершенно определенное, единственное значение второго признака. Такие связи получили название функциональные.

Функциональные связи встречаются в физических и математических обобщениях. Например, площадь треугольника точно определяется его высотой и основанием, длина окружности — радиусом, скорость падения является функцией времени падения и ускорения силы тяжести, скорость протекания определенной химической реакции находится в зависимости от температуры.

Необходимо учесть, что в чистом виде функциональные связи встречаются только в идеальных условиях, когда предполагается, что никаких посторонних влияний нет. На практике это недостижимо. Никогда нельзя точно измерить фактически имеющийся радиус круга, причем вычисленная площадь

никогда неравна в точности фактической, вследствие практической невозможности начертить точную окружность. Скорость падения реального тела в реальных условиях будет всегда различна при одних и тех же времени и ускорении силы тяжести. На практике всегда действуют посторонние для данной функциональной зависимости факторы, которые нарушают точность этой зависимости в разных случаях по-разному.

Пока такие нарушения остаются настолько незначительными, что их практически можно не учитывать, связь считается функциональной.

При изучении живых объектов — диких видов, культурных растений, домашних животных — приходится иметь дело со связями другого рода. Живой организм развивается в связи с условиями его жизни, под действием бесконечно большого числа факторов, которые по-разному определяют развитие разных признаков. У живых объектов связь между любыми двумя признаками настолько часто и сильно нарушается и модифицируется, что не всегда может быть достаточно просто обнаружена.

У растений и животных связь между признаками обычно проявляется особым образом. Тут каждому определенному значению первого признака соответствует не одно значение второго признака, а целое распределение этих значений при вполне определенных основных показателях этого частного распределения — средней величины и степени разнообразия. В этом случае, такая связь называется корреляционной связью или просто корреляцией.

Корреляционная связь, например, между весом животных и их длиной выражается в том, что каждому значению длины соответствует определенное распределение веса (а не одно значение веса), такое, что с увеличением длины увеличивается и средний вес животных. Корреляцию классифицируют по форме и направлению, а измеряют степенью корреляции.

По форме корреляция может быть:

1) прямолинейной;

2) криволинейной.

По направлению:

1) прямо-направленная;

2) обратно-направленная.

Степень корреляции устанавливает силу связи между количественными и качественными признаками. Она измеряется следующими показателями:

1) коэффициентом корреляции r;

2) корреляционным отношением;

3) тетрахорическим и полихорическим показателями связи;

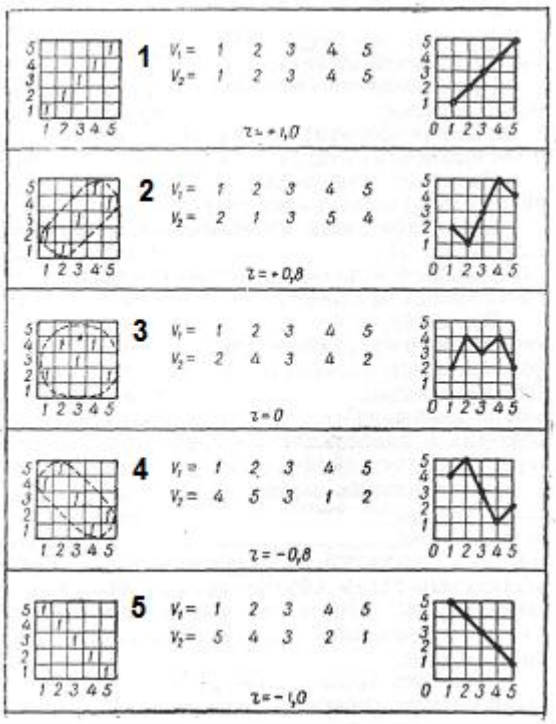
4) частным и множественным коэффициентами корреляции.

Изобразить корреляционную связь двух признаков можно тремя способами (см. рис 5.1), а именно при помощи:

1) корреляционного ряда, состоящего из ряда пар значений признаков;

2) корреляционной решетки, в которой каждой особи соответствует определенная клетка;

3) линии регрессии оси координат, для которой пропорциональны значениям признаков.



1 – прямая полная связь; 2 – прямая частичная связь; 3 – отсутствующая связь; 4 – обратная частичная связь; 5 – обратная полная связь.

Рисунок 5.1 – Корреляционные связи

5.1. Коэффициент корреляции

Коэффициент корреляции измеряет степень и определяет направление прямолинейных связей.

Прямолинейная связь между признаками, это такая связь, при которой равномерным изменениям первого признака соответствуют равномерные (в среднем) изменения второго признака.

Например, при увеличении длины тела на каждый сантиметр, ширина также увеличивается в среднем на 0,7 см.

При графическом изображении прямолинейных связей получается линия, среднее течение которой проходит по прямой.

При измерении степени связи между разными признаками используют их нормированные отклонения, а коэффициент корреляции (r) имеет следующую простую формулу:

$$r = \frac{\sum_{j=1}^{n} x_{1j} \cdot x_{2j}}{\nu}$$

где $x_{ij} = \frac{V_{ij} - M}{\sigma_i}$ – j-е нормированное отклонение i-го признака (i=1; 2);

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{n}(V_{ij} - M)^2}{\nu}}$$

ν – число степеней свободы ν = n - 1;

Сумма произведений нормированных отклонений, входящая в формулу для коэффициента корреляции, обладает следующими тремя особыми свойствами:

1) Если оба признака изменяются параллельно, то сумма произведений их нормированных отклонений дает положительную величину. Если при увеличении одного признака другой уменьшается, то вся сумма будет отрицательной.

Поэтому коэффициент корреляции определяет направление связи: при прямых связях он положителен, а при обратных — отрицателен.

2) При полных связях, когда изменения обоих признаков строго соответствуют друг другу и корреляционная связь превращается в функциональную, сумма произведений нормированных отклонений становится равной числу степеней свободы: $\sum_{j=1}^{n} x_{1j} \cdot x_{2j} = \nu = n - 1$

Поэтому максимальное значение коэффициента корреляции равно единице по абсолютной величине $|r_{\max}| = 1$

для прямых связей: r = +1;

для обратных связей: r = .1.

3) При полном отсутствии корреляционной связи между признаками коэффициент корреляции равен нулю: $r_{min} = 0$.

5.2. Ошибка коэффициента корреляции

Как и всякая выборочная величина, коэффициент корреляции имеет свою ошибку репрезентативности, вычисляемую для больших выборок (n>100) по формуле:

$$m_r = \frac{1 - \left(\bar{r}\right)^2}{\sqrt{n-1}}$$

где $\bar{r}$ — коэффициент корреляции в генеральной совокупности, из которой взята выборка; n — объём выборки, т.е. число пар значений, по которым вычислялся выборочный коэффициент корреляции.

В большинстве исследований значение коэффициента корреляции в генеральной совокупности r неизвестно, поэтому вместо точного значения ошибки коэффициента корреляции берут приближенное значение для выборочного коэффициента корреляции r.

Пример. При исследовании 400 зерен кукурузы найдено, что коэффициент корреляции между длиной и высотой зерна r=+0,85. Определить,

какова возможная величина коэффициента корреляции в генеральной совокупности.

Ошибка найденной величины:

$$m_r = \frac{1-0{,}85^2}{\sqrt{399}} = 0{,}014$$

Отсюда при t_1=2,0 генеральный коэффициент корреляции:

$$\bar{r} = r \pm t \cdot m_r = 0{,}85 \pm 2 \cdot 0{,}014$$

Для малых выборок (n<100) необходимо пользоваться другой формулой ошибки выборочного коэффициента корреляции:

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

Критерий достоверности коэффициента корреляции, определяемый по формуле:

$$t_r = \frac{r}{m_r} \geq t_{st}$$

оценивается путем сравнения фактически полученного значения tr с фиксированными значениями tst (стандартное значение критерия Стьюдента), которые соответствуют трем степеням вероятности безошибочных прогнозов. При этом для t_{st} степень свободы равна $\nu = n - 2$.

- если расчетное $t_r > t_{st}$, то коэффициент корреляции достоверен, и можно считать, что между исследуемыми признаками существует взаимосвязь.
- если $t_r < t_{st}$, то коэффициент корреляции недостоверен, и нельзя сделать вывод о взаимосвязи между исследуемыми признаками в выборке, а также в генеральной совокупности.

Пример. Для выяснения силы действия модифицирующих факторов (плодородия почвы, климатических условий) при сравнении двух сортов кукурузы взяты 20 соседних участков, на которых попарно были высеяны один и другой сравниваемые сорта, а затем рассчитан коэффициент корреляции между урожаями сравниваемых сортов.

Большой коэффициент должен указывать на слабое действие модифицирующих агентов, а малый – должен означать, что различия между парными участками по урожаю подверглись в течение опыта каким-то сильным и разнообразным влияниям. В итоге были получены следующие результаты наблюдений:

объем выборки n=20;

коэффициент корреляции r=+0,63.

Определим достоверность коэффициента корреляции:

$$m_r = \sqrt{\frac{1-0{,}63^2}{20-2}} = 0{,}18\,; \qquad t_r = \frac{0{,}63}{0{,}18} = 3{,}5$$

ti={2,1 (b1=0,95); 2,85 (b2=0,99); 3,85 (b3=0,999)}

Таким образом, получен достоверный коэффициент корреляции между урожаями соседних участков, для порога вероятности не выше 0,999. Это означает, что различия опытных участков по плодородию почвы и другим факторам, определяющим урожай, были слабы и не дали проявиться различию испытуемых сортов.

5.3. Частный коэффициент корреляции

В некоторых исследованиях требуется выяснить, не является ли связь между двумя признаками обусловленной влиянием какого-нибудь третьего признака. Например, при изучении статистических связей между урожаем и средней температурой воздуха имеет смысл учесть, влияние третьего признака — количества осадков, который влияет на оба признака — и на урожай, и на среднюю температуру воздуха. Для того, чтобы выяснить в таких исследованиях, влияние третьего признака на корреляционную связь между первым и вторым признаком, необходимо исследовать эту связь при его постоянном значении.

При постоянном значении признака можно только констатировать, что в изменчивости других признаков нет его влияния: он постоянен, а другие признаки изменяются.

Коэффициент корреляции между первым и вторым признаками при постоянном значении третьего признака называется частным коэффициентом корреляции и обозначается символом $r_{12\cdot3}$.

Для его расчета не всегда нужно проводить рассмотренный выше эксперимент. Если связь между парой признаков прямолинейна или отличается от прямолинейной незначительно, то величину частного коэффициента корреляции можно определить по обычным коэффициентам корреляции:

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\cdot r_{23}}{\sqrt{\left(1-r_{13}^2\right)\cdot\left(1-r_{23}^2\right)}}, \text{ при } 3 = \text{const}$$

где: r_{12} – коэффициент корреляции между 1 и 2 признаком;

r_{13} – коэффициент корреляции между 1 и 3 признаком;

r_{23} – коэффициент корреляции между 2 и 3 признаком.

Пример. При исследовании корреляционной связи между весом животных (признак 1) и диаметром мускульных волокон (признак 2), без влияния на эту связь калорийности пищи (признак 3), (т.е. при постоянном значении калорийности пищи) были получены следующие коэффициенты корреляции: между весом и диаметром волокон r_{12}=+0,6 (без выравнивания калорийности пищи);

между весом и калорийностью r_{13}=+0,8;

между диаметром волокон и калорийностью r_{23}=+0,7.

Частный коэффициент корреляции:

$$r_{12\cdot3} = \frac{0,6 - 0,8 \cdot 0,7}{\sqrt{(1-0,8^2)\cdot(1-0,7^2)}} = 0,09\,; \qquad r_{13\cdot2} = 0,67 \qquad r_{23\cdot1} = 0,46$$

Выявилась очень малая частная корреляция. Исследование показало, что если исключить статистическое влияние калорийности пищи, т.е. выровнять калорийность рационов, то между весом животных и диаметром их мускульных волокон не будет почти никакой корреляции, хотя обычно, без выравнивания калорийности пищи, эта связь внешне выражается довольно значительным коэффициентом: +0,6.

При совместном изучении трех признаков можно исключить влияние не только третьего, но также и первого или второго признака:

$$r_{13\cdot2} = \frac{r_{13} - r_{12} \cdot r_{23}}{\sqrt{(1-r_{12}^2)\cdot(1-r_{23}^2)}}\text{, при 2 = const}$$

Иногда вычисление частного коэффициента корреляции дает результаты, кажущиеся на первый взгляд невероятными. Однако, при более внимательном анализе явления, уже не с математической, а со специальной точки зрения эти результаты становятся вполне понятными и легко объяснимыми.

Пример. При изучении зависимости веса древесины (3) от размеров дерева: обхвата (длина периметра сечения) на уровне груди (1) измеряющего и высоты (2) ствола — были получены следующие коэффициенты корреляции:

между обхватом (1) и высотой (2): r_{12}= +0,5;

между обхватом (1) и весом (3): r_{13}= + 0,9 ;

между высотой (2) и весом (3):. r_{23}= + 0,8

Частные коэффициенты корреляции каждого размера с весом при исключенном влиянии другого размера не вызывают никаких недоумений и указывают на большую частную корреляцию обхвата и высоты с весом древесины:

$$r_{13\cdot2} = \frac{0,90 - 0,50 \cdot 0,80}{\sqrt{(1-0,25)\cdot(1-0,64)}} = +0,96$$

$$r_{23\cdot1} = \frac{0,80 - 0,50 \cdot 0,90}{\sqrt{(1-0,25)\cdot(1-0,81)}} = +0,92$$

Частная корреляция между обоими размерами при исключенном влиянии веса, т. е. при его постоянном значении:

$$r_{12\cdot3} = \frac{0,50 - 0,90 \cdot 0,80}{\sqrt{(1-0,81)\cdot(1-0,64)}} = -0,84$$

Оказалось, что между обхватом и высотой дерева получилась значительная отрицательная частная зависимость: при увеличении высоты, обхват дерева уменьшается. Это, казалось бы, явно противоречит обычным процессам развития деревьев: если увеличивается высота, то, конечно, увеличивается и обхват.

Объяснение этого мнимого противоречия заключается в основном условии частной корреляции — постоянстве исключаемого признака.

Если взять деревья одного и того же веса, то среди таких деревьев увеличение высоты может происходить только за счет уменьшения обхвата. Если бы увеличивались оба размера, то вес древесины не мог бы оставаться постоянным.

При корреляции 4-х признаков, расчет ведут по формуле:

$$r_{12\cdot34} = \frac{r_{12\cdot4} - r_{13\cdot4} \cdot r_{23\cdot4}}{\sqrt{\left(1 - r_{13\cdot4}^2\right) \cdot \left(1 - r_{23\cdot4}^2\right)}}$$

Множественный коэффициент корреляции показывает взаимосвязь между всеми изучаемыми признаками одновременно.

5.4. Ошибка частного коэффициента корреляции

Ошибка репрезентативности выборочного частного коэффициента корреляции рассчитывается по такой же формуле, как и в случае обычного коэффициента корреляции для малочисленны групп (n<100):

$$m_{r_{12}} = \sqrt{\frac{1 - r^2}{n - 2}}$$

При оценке критерия достоверности частного коэффициента корреляции, предельные значения показателя вероятности берутся для числа степеней свободы, которые соответствуют v=n-2-k, где k – число элиминированных признаков.

Объем выборки в данном случае равен числу пар значений (n), одинаковому для всех обычных коэффициентов корреляции, которые необходимы для расчета частного коэффициента корреляции.

5.5. Коэффициент прямолинейной регрессии

Как уже упоминалось, прямолинейная корреляция отличается тем, что при этой форме связи каждому из одинаковых изменений первого признака соответствует вполне определенное и тоже одинаковое в среднем изменение другого признака, связанного с первым или зависящего от первого.

Та величина, на которую в среднем изменяется второй признак, при изменении первого на единицу измерения, называется коэффициентом прямолинейной регрессии. Рассчитывается он по следующей формуле:

$$R = \frac{\sigma_2}{\sigma_1} \cdot r_{12},$$

где R – коэффициент прямолинейной регрессии;

σ_2 – среднее квадратическое отклонение второго признака, который изменяется в связи с изменением первого;

σ_1 – среднее квадратическое отклонение первого признака, в связи с изменением которого изменяется второй признак;

r_{12} – коэффициент корреляции между первым и вторым признаками.

Ошибка коэффициента регрессии равна ошибке коэффициента корреляции, умноженной на отношение сигм:

$$m_R = \frac{\sigma_2}{\sigma_1} \cdot m_r = \frac{\sigma_2}{\sigma_1} \cdot \sqrt{\frac{1-r^2}{n-2}}$$

Критерий достоверности коэффициента регрессии равен критерию достоверности коэффициента корреляции:

$$t_R = \frac{R}{m_R} = \frac{\frac{\sigma_2}{\sigma_1} \cdot r_{12}}{\frac{\sigma_2}{\sigma_1} \cdot m_r} = \frac{r}{m_r} = t_r$$

Пример. При разработке методов селекции молочного скота выяснялась связь высшего суточного удоя с удоем за 300 дней той же лактации. Всего изучено 577 лактаций, проходивших в оптимальных условиях. Были получены следующие данные:

высший суточный удой (признак 1): $n_1 = 577$; $M_1 = 17{,}2$ кг; $\sigma_1 = 3{,}9$ кг;

удой за 300 дней лактации (признак 2): $n_2 = 577$; $M_2 = 3250$ кг; $\sigma_2 = 685$ кг;

коэффициент корреляции: $r_{12} = +0{,}829$.

Рассчитаем коэффициент регрессии удоя за 300 дней по высшему суточному удою:

$$\tilde{R} = \frac{685}{3{,}9} \cdot (+0{,}829) = +145{,}6 \text{ кг} \qquad m_R = \frac{685}{3{,}9} \cdot \sqrt{\frac{1 \cdot 0{,}829^2}{575}} = 4{,}2$$

$$\overline{R} = \tilde{R} \pm 2m_R = +145{,}6 \pm 2 \cdot 4{,}2 \quad \begin{cases} \text{не менее} + 137{,}2 \text{ кг} \\ \text{не более} + 153{,}8 \text{ кг} \end{cases}$$

Вычисления показали, что в данном случае генеральный коэффициент регрессии равен R = +145,6 ± 4,2 кг. Это значит, что при увеличении высшего суточного удоя на каждый 1 кг удой за 300 дней лактации увеличивается на +145,6 кг с возможными отклонениями этой величины в пределах 138 – 154 кг.

Таким образом, если, например, у группы коров высший суточный удой в среднем на 5 кг больше среднего по сверстницам, то можно ожидать, что удой за 300 дней лактации этих коров будет на 5•145,6=728 кг больше среднего по их сверстницам.

5.6. Тетрахорический показатель связи

При альтернативном разнообразии, когда оба качественных признака выражаются только наличием или отсутствием их у особей, корреляционная связь между двумя признаками измеряется тетрахорическим показателем связи.

Если у каждой особи изучаются два признака, то вся группа разбивается на следующие четыре части:

[a] — особи, имеющие оба признака (+ +);

[b] — особи, имеющие первый признак, но не имеющие второго (+ —);

[c] — особи, не имеющие первого признака, но имеющие второй (— +);

[d] — особи, не имеющие обоих признаков (— —).

Если обозначить численность указанных четырех групп этими же буквами (a, b, c, d), то степень связи наличия первого признака с наличием второго признака будет определяться тетрахорическим показателем связи, который вычисляется по следующей формуле:

$$r_{++} = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (a+c) \cdot (d+b) \cdot (d+c)}}$$

Пример. При проверке эффективности действия прививки против сыпного тифа получены первичные материалы о числе заболевших (—) и не заболевших (+) из числа получавших (+) и не получавших (—) прививку (см. табл.5.1). Объём выборки n=210 человек.

Таблица 5.1

Признак 2	Признак 1		Σ
	Получили прививку (+)	Не получили прививку (-)	
Не заболели (+)	(a)++=54	(c)-+=106	(a+c)=160
Заболели (-)	(b)+-=6	(d)--=44	(b+d)=50
Σ	(a+b)=60	(c+d)=150	N=210

$$r_{++} = \frac{54 \cdot 44 - 6 \cdot 106}{\sqrt{60 \cdot 150 \cdot 160 \cdot 50}} = \frac{+1740}{8485{,}3} = +0{,}205$$

Достоверность тетрахорического коэффициента корреляции определяется по величине χ^2_{r++} (кси).

$$\chi^2_{r++} = n \cdot r^2_{++}, \qquad \chi^2_{r++} \geq \chi^2_{st}$$

n – общий объем выборки.

Возвращаясь, к примеру, предположим n=210, тогда:

$$\chi^2_{r++} = 210 \cdot 0{,}205^2 = 8{,}8$$

Величине χ^2_{r++} должно соответствовать табличное значение, которое зависит от ответственности порога вероятности (β). Число степеней свободы для тетрахорического коэффициента равно 1, т.е. число градаций минус единица.

Таблица 5.2 – Стандартные значения χ^2_{st}

β	0,95	0,99	0,999
χ^2_{st}	3,8	6,6	10,8

5.7. Полихорический показатель связи

Существуют такие количественные признаки, степень развития которых характеризуется не результатом точного измерения, не числом, а качественными градациями, которые определяются субъективно, путем осмотра или вкусовой пробы.

Например:

а) цвет пера птицы – светло-серый, серый и темно-серый;

б) вкус сливочного масла – слабо-, средне- и сильносоленый;

в) упитанность животных – жирная, вышесредняя, средняя, нижесредняя, тощая и т.д.

Определение степени корреляционной связи между такими признаками можно производить при помощи полихорического показателя связи, обозначаемого греческой буквой . и вычисляемого по следующей формуле:

$$\rho = \frac{a-1}{\sqrt{(gr_1-1)\cdot(gr_2-1)}}, \qquad \text{где:} \quad a = \sum\left(\frac{\sum\frac{f^2}{n_2}}{n_1}\right);$$

f – частоты ячеек корреляционной решетки по первому и второму признакам;

n1 — частоты ряда первого признака (определяются по столбцам в нижней суммарной строке корреляционной таблицы-решетки);

n2 — частоты ряда второго признака (определяются по строкам в правом суммарном столбце корреляционной таблицы-решетки);

gr1, gr2 — число градаций, на которые разбиты первый и второй признаки;

n — общая численность группы: $n=\Sigma n_1=\Sigma n_2$.

Полихорический показатель связи всегда выражается положительным числом, поэтому определение характера связи (прямая/обратная) производится по виду корреляционной решетки.

Пример. При исследовании связи между крепостью телосложения одного вида животных (признак 1) и густотой их шерсти (признак 2) получены следующие данные (см. табл.5.3).

Таблица 5.3 – Зависимость признака 1 от признака 2

Шерсть (признак 2)	Плотность телосложения (признак 1)			n_2
	Сильная	Средняя	Слабая	
Густая	30	9	1	40
Средняя	5	21	4	30
Редкая	2	3	25	30
n_1	37	33	30	100

Расчет полихорического показателя связи нужно производить по следующим этапам:

1. Подсчитать частоты n_1 по первому признаку – суммы по столбцам (37, 33, 30), затем частоты n_1 по второму – суммы по строкам и общую численность группы $n=n_1=n_2=37+33+30=40+30+30=100$.

2. В каждой ячейке возвести в квадрат частоту и полученный результат f^2 записать в той же ячейке в скобках. Затем квадрат частоты ячейки разделить на частоту второго признака по той же строке, в которой находится ячейка, и полученный результат $\frac{f^2}{n_2}$ f записать в той же ячейке под ранее записанной цифрой (см. табл.5.4).

3. Последние числа ячеек $\frac{f^2}{n_2}$ сложить по столбцам, т.е. по градациям второго признака.

4. Полученные значения разделить на частоты ряда первого признака n_1.

5. Найти сумму значений цифр последней строки. Это будет величина a.

6. Значения a, gr_1, gr_2, n подставить в формулу для полихорического показателя связи.

Сведение материалов в корреляционную решетку выявило вполне заметную связь между изучаемыми признаками: при сильной крепости телосложения большинство особей имело густую шерсть. Степень связи между этими признаками определим, рассчитав полихорический показатель связи:

$$gr_1 = 3; \qquad gr_2 = 3; \qquad n=100; \qquad \rho = \frac{1{,}86-1}{\sqrt{(3-1)\cdot(3-1)}} = 0{,}43$$

Между изученными признаками имеется корреляционная связь = 0,43.

Таблица 5.4 – Результаты расчёта

Шерсть (признак 1)	Плотность телосложения (признак 2)			n_2
	Сильная	Средняя	Слабая	
Густая	f^2=(900) f^2/n_2=22,5	(81) 2,02	(1) 0,02	40
Средняя	(25) 0,83	(441) 14,7	(16) 0,53	30
Редкая	(4) 0,13	(9) 0,30	(625) 20,83	30
n_1	37	33	30	100
$\Sigma(f^2:n_2)$	23,46	17,02	21,38	-
$\frac{\Sigma(f^2:n_2)}{n_1}$	0,63	0,52	0,71	a=1,86

Достоверность полихорического показателя связи можно определить при помощи критерия χ^2, который для данного показателя равен $\chi = n\cdot(a-1) \geq \chi^2_{st}$ при числе степеней свободы $\nu=(gr_1-1)\cdot(gr_2-1)$. Для нашего примера:

$$\chi^2 = 100 \cdot (1{,}86 - 1) = 86{,}0 \qquad \chi^2_{st} = (18{,}5 \;\; 13{,}3 \;\; 9{,}5) \;\; \text{для } \nu = 2 \cdot 2 = 4$$

$$\chi^2 > \chi^2_{st}$$

5.8. Проверка артефактов (выпадов)

Артефактом (выпадом) – называется резко выделяющееся значение признака (очень большое или маленькое) на фоне остальных значений этого признака.

Проверка артефактов должна быть проведена перед началом обработки любых экспериментальных данных. Если подтвердится, что сильно выпадающие (выделяющиеся) значения не могут относиться к объектам данной группы и попали в записи наблюдений из-за ошибок внимания, то такой артефакт исключается из обработки. Проверка на артефакт проводится исходя из следующего условия:

$$T = \frac{\left|\hat{V} - M\right|}{\sigma} \geq T_{st},$$

T – критерий артефакта; $\hat{V}$ – выделяющееся значение признака (возможный артефакт); T_{st} – табличное значение критерия артефакта; σ - среднеквадратичное отклонение (без артефакта); M – среднее значение (без артефакта).

Стандартные значения критерия выпадов сведены в таблицу 5.5

Таблица 5.5 - Стандартные значения критерия выпадов

N	2	3-4	4-9	10-15	16-20	21-28	29-34	35-46
T_{st}	2	2,1	2,2	2,3	2,4	2,5	2,6	2,7
N	47-66	67-84	85-104	105-124	125-174	175-349	350-599	600-1500
T_{st}	2,8	2,9	3	3,1	3,2	3,3	3,4	3,5

Пример. Получены значения признака: 1, 2, 3, 10. Определить является ли число 10 в этой последовательности артефактом.

n=3, M=3,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(V_i - M)^2}{n-1}} = \sqrt{\frac{(1-2)^2 + (2+2)^2 + (3-2)^2}{3-1}} = 1$$

$$T = \frac{10-2}{1} = 8 \geq 2{,}1$$

Таким образом, число 10 выпадает и его можно исключить из рассмотрения.

6. ДИСПЕРСИОННЫЙ АНАЛИЗ

Для того, чтобы, пользуясь дисперсионным анализом, получить правильные результаты, необходимо выполнять определенные правила организации дисперсионных комплексов. Вначале рассмотрим несколько важных понятий дисперсионного анализа.

Дисперсионный анализ заключается в изучении статистического влияния одного или нескольких факторов на результативный признак (y).

Результативный признак (y) – это признак, который изучается как результат статистического влияния факторов: организованных (контролируемых или учитываемых) (x) и всех остальных неорганизованных в данном исследовании (z).

Результативными признаками могут быть:

- точно измеряемые количественные особенности объектов – длина, ширина, резвость, шерстность;
- неточно измеряемые особенности – густота крема, цвет, умственные способности;
- комбинированные признаки – соотношения размеров тела, индексы продуктивности;
- качественные признаки – масть, болезнь, выздоровление.
- отдельные признаки, принимаемые за аргумент при изучении среднего признака, принимаемого за функцию.

Фактор – это любое влияние, воздействие или состояние, разнообразие которых может отражаться на разнообразии результативного признака.

Факторы могут быть:

- физического влияния – температура, влажность, давление;
- химического влияния – питание;
- биологического влияния – наличие мутагенов, возраст, пол, сорт, национальность.

Градациями факторов называется степень их действия или состояния объектов изучения. В качестве градации факторов могут выступать разная температура, влажность, доза облучения, различная продолжительность воздействия, различная питательность и состав корма, дозы стимулирующих и химических мутагенов, разные периоды болезни, степень таланта, разные родители, разные ареалы обитания, разные условия жизни.

Градациями комплекса называются опытные группы исследований. Каждая градация комплекса соответствует одной градации фактора и включает те объекты с их датами, которые подвергаются одной степени воздействия фактора или находится в одном из изучаемых состояний. Организация градаций комплекса осуществляется различными способами, такими как подбор опытных и контрольных групп, использование ранее полученных результатов исследования, систематизация записей производственной отчетности.

Обычно комплекс представляют в виде таблицы, столбцы которой соответствуют градациям (комбинациям) факторов.

Разнообразие – это наличие неодинаковых значений каждого признака у разных особей объединенных в одну группу. Как отмечалось ранее, разнообразие группы особей по изучаемому признаку измеряется показателями разнообразия: лимитами, средним квадратическим отклонением, коэффициентами вариации.

Для того, чтобы выяснить степень и достоверность влияния изучаемых организованных и неорганизованных факторов, измеряют ту часть общего разнообразия, которая вызывается этими факторами. Делается это при помощи двух величин: дисперсии и девиаты.

Дисперсия – это первичная мера разнообразия в рассматриваемой группе. Она равна сумме квадратов центральных отклонений. Общая дисперсия признака определяется как:

$$C_y = \sum_{i=1}^{n} (V_i - M)^2$$

Общее разнообразие результативного признака всегда больше того разнообразия, которое связано со статистическим влиянием организованных факторов. Происходит это потому, что в любом исследовании нельзя освободиться от действия всего множества остальных факторов, так или иначе влияющих на изменение результативного признака.

Поэтому при проведении дисперсионного анализа общая дисперсия признака C_y в изучаемой группе расчленяется на две дисперсии - факториальную или частную (вызванную организованными факторами) C_x, и случайную или остаточную дисперсию (вызванную остальными, неорганизованными в данном опыте факторами) C_z. Сумма факториальной и случайной дисперсий всегда равна общей:

$$C_y = C_x + C_z$$

Факториальная дисперсия:

$$C_x = \sum (M_x - M)^2$$

Случайная дисперсия:

$$C_z = \sum (V - M_x)^2$$

где M_x — частная средняя результативного признака по каждой отдельной градации организованных факторов.

Дисперсия как показатель разнообразия зависит от числа особей в группе. Для определения степени влияния факторов это обстоятельство не имеет значения. Для других же целей, в частности для установления достоверности влияния факторов, обнаруженного при выборочном исследовании, необходим показатель, свободный от указанной зависимости, допускающий сравнение групп, различных по числу входящих в них элементов. Таким показателем является девиата.

Девиатой называют дисперсию, приходящуюся на один элемент свободного разнообразия или на одну степень свободы:

$$\sigma_y^2 = \frac{C_y}{\nu_y} \qquad \sigma_x^2 = \frac{C_x}{\nu_x} \qquad \sigma_z^2 = \frac{C_z}{\nu_z}$$

где σ_y^2 — общая девиата по всему комплексу; σ_x^2— факториальная девиата; σ_y^2 — случайная девиата; ν — число степеней свободы.

Корень квадратный из девиаты является средним квадратическим отклонением:

$$\sigma = \sqrt{\sigma_y^2}$$

Девиаты используются в дисперсионном анализе для определения достоверности влияния организованных факторов на результативный признак, обнаруженного в выборочном исследовании. Достоверность влияния организованного фактора определяется отношением факториальной девиаты к случайной, которое должно быть не меньше табличной стандартной величины F_{st}:

$$F = \frac{\sigma_x^2}{\sigma_z^2} \geq F_{st}$$

Если это отношение равно или больше определенной стандартной величины F_{st}, влияние считается достоверным с определенной степенью вероятности. Стандартные отношения девиат F_{st} определяются по специальным таблицам.

Пример. Изучается действие на рост растений (у), который является результативным признаком двух факторов: А (температура среды) и В (влажность). Каждый фактор берется в двух градациях: A_1 и A_2 - низкая и высокая температура; B_1 и B_2 малая и большая влажность.

Для каждой из четырех градаций двух факторов — A_1B_1, A_1B_2, A_2B_1, A_2B_2 — по способу случайной выборки выбрано по две особи. У всех восьми особей измерен результативный признак и результаты записаны в виде статистического комплекса в табл.6.1.

Таблица 6.1

Градации 1-го фактора	A_1				A_2			
Градации 2-го фактора	B_1		B_2		B_1		B_2	
Значения результативного признака	9	11	3	5	1	3	7	9

Необходимо проанализировать полученный комплекс и установить следующее: оказывают ли влияние на результативный признак – рост, изучаемые факторы – температура и влажность в их общем суммарном действии. какова роль каждого фактора в отдельности и в их сочетаниях?

Для решения, вначале допустим, что действуют не два, а один суммарный фактор х, имеющий все указанные 4-ре градации, которые организованы в исследовании для всех рассматриваемых факторов.

Найдем общую дисперсию Су. Для этого рассчитаем общую среднюю:

$$M = \frac{9+11+3+5+1+3+7+9}{8} = 6$$

тогда: $C_y=(9-6)^2+(11-6)^2+(3-6)^2+(5-6)^2+(1-6)^2+(3-6)^2+(7-6)^2+(9-6)^2=88$

Найдем факториальную дисперсию. Расчет сведем в табл.6.2.

Таблица 6.2 – Результаты расчёта

Градации факторов	x_1		x_2		x_3		x_4		Итог
	x_{11}	x_{12}	x_{21}	x_{22}	x_{31}	x_{32}	x_{41}	x_{42}	
Значения результативного признака	9	11	3	5	1	3	7	9	n=8
Сумма значений	20		8		4		16		Σ=48
Частные средние M_x	$M_{x1}=10$		$M_{x2}=4$		$M_{x3}=2$		$M_{x4}=8$		M=6
Центральные отклонения средних $(M_x - M)$	+4		-2		-4		+2		
$(M_x - M)^2$	16	16	4	4	16	16	4	4	$C_x=80$
Частные центральные отклонения $(V- M_x)$	-1	+1	-1	+1	-1	+1	-1	+1	
$(V- M_x)^2$	1	1	1	1	1	1	1	1	$C_z=8$

Количество квадратов центральных отклонений дат берут столько, сколько дат результативного признака (=8).

На рис. 6.1 представлена зависимость изменения частной средней от градации факторов.

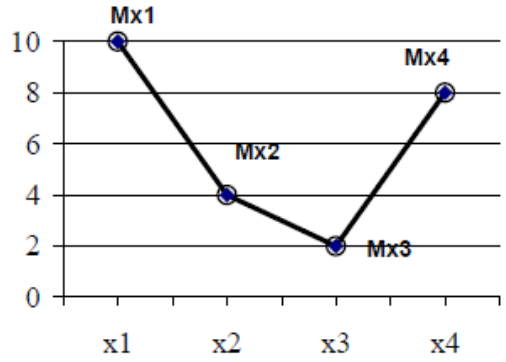


Рисунок 6.1 – Изменение частной средней от градации факторов

Степень влияния фактора на результативный признак равно отношению факториальной дисперсии к общей:

$$\eta_x^2 = \frac{C_x}{C_y} = \frac{80}{88} = 0{,}91$$

Полученный результат означает, что 91% всего разнообразия результативного признака определяется разнообразием организованных факторов.

Найдем случайную дисперсию. Расчет сведен в табл.6.2. Степень влияния неорганизованных факторов:

$$\eta_x^2 = \frac{C_z}{C_y} = \frac{8}{88} = 0{,}09$$

Таким образом, влияние неорганизованных факторов составляет всего 9% от общего влияния всех факторов. Это указывает на большую силу влияния суммарного фактора в виде температуры и влажности на результативный признак – рост.

График показывает, что при изменении фактора от x1 до x4 результативный признак сначала уменьшается, а потом возрастает. Такое влияние одного не объединенного фактора встречается редко, но в нашем случае градациями факторами x являются четыре комбинации градаций двух факторов.

Определим достоверность влияния организованного фактора с помощью девиаты.

а) для общей дисперсии число степеней свободы равно:

$$\nu = n-1=8-1=7,$$

где n – число градаций результативного признака (y);

б) для факториальной дисперсии число степеней свободы равно:

$$\nu_x = r_x-1=4-1=3,$$

где r_x – число градаций организованного суммарного фактора;

факториальная девиата:

$$\sigma_x^2 = \frac{C_x}{\nu_x} = \frac{80}{3} = 26{,}7$$

в) для случайной дисперсии число степеней свободы равно:

$$\nu_z = n - r_x = 8-4 = 4$$

случайная девиата:

$$\sigma_x^{\ 2} = \frac{C_z}{\nu_z} = \frac{8}{2} = 2{,}0$$

Достоверность влияния: $F = \frac{\sigma_x^2}{\sigma_z^2} = \frac{26{,}7}{2} = 13{,}4 \geq F_{st} = 6{,}6(b_1 = 0{,}95)$

для b2 = 0,99 – Fst = 16,7

для b3 = 0,999 – Fst = 56,1

Это значит, что наблюдаемое влияние организованного суммарного фактора с достоверностью b1=0,95 не является случайным, так как факториальная дисперсия оказалась большей чем случайная.

Таким образом, суммарное действие двух факторов A и B на результативный признак очень велико и достоверно. Остается выяснить:

1) каково значение каждого из этих факторов в отдельности при выровненном действии другого.

2) каково значение различий их совместного действия при разных комбинациях градаций.

Для поиска ответов на эти вопросы используют развернутый комплекс, который называется двухфакторным.

Решение двухфакторного комплекса на рассматриваемом примере проводится по следующим этапам:

1. выполняют определение общей дисперсии – аналогично однофакторному комплексу:

$$C_y=88$$

2. выполняют определение случайной дисперсии Cz – аналогично однофакторному комплексу:

частные средние Mx по 4-м градациям: 10, 4, 2, 8;

отклонения от своей частной средней (V – Mx): -1, +1, -1, +1, -1, +1, -1, +1;

квадраты отклонений $(V - Mx)^2$: +1, +1, +1, +1, +1, +1, +1, +1;

их сумма $\Sigma(V - M_x)^2$: $C_z=8$.

3. Определение дисперсии суммарного действия организованных факторов аналогично однофакторному комплексу (см. табл.6.2): $C_x = 80$.

4. Определение частных факториальных дисперсий отдельно по каждому фактору: расчет частных средних MA и дисперсий по фактору A:

$$C_A=\Sigma(M_A-M)$$

расчет частных средних MB и дисперсий по фактору B:

$$C_B=\Sigma(M_B-M)$$

Результаты сведены в табл. 6.3

5. Степень влияния каждого фактора:

$$\eta_A^2 = \frac{C_A}{C_y} = \frac{8}{88} = 0{,}09 \qquad \eta_B^2 = \frac{C_B}{C_y} = \frac{0}{88} = 0$$

Полученный результат свидетельствует о том, что влияние фактора B при выровненных значениях фактора A не проявляется в разнообразии результативного признака.

Такое соотношение показателей η_A^2 = 9 % , η_B^2 = 0 %, η_x^2 = 91 % отражает влияние одних факторов на другие и показывает, что при нормальной температуре A_1 нормальная влажность B_1 благоприятна для роста, а повышенная влажность B_2 уже угнетает рост, т.е. при сочетании градаций A_1B_1: $M_x=10$, а при сочетании A_1B_2: $M_x=4$.

При повышенной температуре A_2, наоборот, низкая влажность B1 недостаточна для нормального роста (сочетание A_2B_1: $M_x=2$), и по этому он замедлен, а повышенная влажность B_2 благоприятна, и рост усиливается (сочетание A_2B_1: $M_x=8$).

Таблица 6.3 - Дисперсионный анализ двухфакторного комплекса

n=8	Градации 1-го фактора	A_1				A_2			
ΣV=48	Градации 2-го фактора	B_1		B_2		B_1		B_2	
	Значения результативного признака (рост)	9	11	3	5	1	3	7	9
M=6	M_x	$M_{x1}=10$		$M_{x2}=4$		$M_{x3}=2$		$M_{x4}=8$	
	(M_x-M)	+4		-2		-4		+2	
$C_x=80$	$(M_x-M)^2$	16	16	4	4	16	16	4	4
Градации фактора А		A_1				A_2			
	Значения по градациям	9	11	3	5	1	3	7	9
	Сумма по градациям А	28				20			
	Частные средние M_A	$\frac{28}{4}=7$				$\frac{20}{4}=5$			
	(M_A-M)	7-6=+1				5-6=-1			
	$(M_A-M)^2$	1	1	1	1	1	1	1	1
Градации фактора В		B_1				B_2			
	Значения по градациям	9	11	1	3	3	5	7	9
	Сумма по градациям В	24				24			
	Частные средние M_B	$\frac{24}{4}=6$				$\frac{24}{4}=6$			
M=6	(M_B-M)	6-6=0				6-6=0			
$C_B=0$	$(M_A-M)^2$	0	0	0	0	0	0	0	0

Если рассматривать каждый фактор в отдельности, то температура без регулирования влажности и влажность без регулирования температуры сами по себе слабо проявляются в разнообразии роста: η_A^2= 9 % , η_B^2= 0 %.

Если организовать определенное сочетание факторов (их градаций), например, при определенной температуре обеспечить определенную влажность, то различные комбинации этих факторов создадут значительное разнообразие результативного признака (роста растений), что и покажет их большое суммарное действие: η_x^2 = 91% x .Поскольку всегда имеется некоторое различие в действии одного фактора при различных градациях другого, то в каждой градации дисперсионного комплекса суммарное действие всех организованных факторов слагается из действия каждого фактора в отдельности и специфического действия от их сочетаний.

6.1. Подбор факторов для дисперсионного анализа

При организации однофакторных комплексов фактором считается любой признак, влияние которого на результативный признак требуется изучить. Это могут быть другие признаки того же животного или растения, различные условия жизни, химические или биологические агенты и другие влияния.

При организации двух- и многофакторных комплексов свободный выбор факторов для исследования ограничен требованием полной независимости их между собою. Для таких комплексов нельзя в качестве двух факторов брать, например, вес и размер животных, так как эти признаки нельзя подбирать независимо друг от друга: при малом весе невозможно подобрать такие же значения размера, как и при большом весе.

Независимыми факторами могут быть, например, температура а влажность, пол и уровень кормления, химическое и биологическое воздействие.

6.2. Разделение факторов на градации

При проведении дисперсионного анализа не требуется, чтобы факторы были разделены обязательно на количественные градации в форме вариационного ряда. Как для однофакторных, так и для двух- и многофакторных комплексов факторы могут иметь и качественные градации, например, пол — мужской, женский; цвет волоса или пера светло-серый, серый, темно-серый; упитанность — жирная, выше средней, средняя, ниже средней; крепость телосложения — слабая, нормальная, сильная.

При установлении градации факторов нужно помнить, что результаты дисперсионного анализа в большой степени зависят от того уровня, на котором установлены градации факторов.

Если, например, изучается действие температуры, то при градациях 15°, 20°, 25°С может быть найдено достоверное влияние этого фактора на результативный признак, но это совсем не значит, что такое же сильное влияние будет при другом уровне градаций, например, 5°, 10°, 15°С.

Большое значение также имеет уровень группы неорганизованных факторов, которые составляют фон дисперсионного анализа. Например, комбинированное действие возраста и какого-нибудь стимулятора ожирения дают при одном уровне общего кормления и содержания определенный эффект, а при другом, например скудном кормлении и плохом содержании, может и совсем не проявиться.

6.3. Подбор особей. Типы комплексов

Результаты дисперсионного анализа в основном зависят от того, насколько правильно подобраны особи, как по качеству, так и по количеству. По своему качеству особи для дисперсионного анализа должны отражать ту генеральную совокупность, для изучения которой и проводится исследование.

По величине результативного признака особи должны быть подобраны по принципу случайной выборки. Лучше всего при отборе объектов для дисперсионного анализа поступать следующим образом.

Пусть для данной градации требуется, например, 20 особей, а всего имеется 30 особей. Тогда номер каждой особи нужно записать на карточку. Все 30 карточек хорошо перетасовать и взять подряд без выбора первые 20

карточек или, наоборот, взять подряд только первые 10 карточек. В первом случае отберутся особи для исследования, во втором откинутся особи, лишние для данной градации.

Организация дисперсионного комплекса с выполнением принципа случайности отбора вариантов называется рандомизацией, а комплексы, организованные таким образом, называются рандомизированными.

По количеству особи могут распределяться по градациям факторов различными способами: поровну, пропорционально, неравномерно. В соответствии с этим организованные комплексы бывают равномерными (подбирается одинаковое число дат) и неравномерными (подбирается не одинаковое число дат).

Если градации двух или многофакторных комплексов заполнены разным числом дат, но таким образом, что даты по градации одного фактора находятся в одинаковом отношении (пропорции) ко всем остальным факторам, то комплекс называется пропорциональным (табл 6.4). Равномерные и пропорциональные комплексы называются ортогональными. Равномерный комплекс является частным случае пропорционального, когда отношение частот равно 1:1; 1:1; 1:1 и т.д. (см. табл.6.5).

Таблица 6.4 – Пропорциональный комплекс

A_1		A_2	
B_1	B_2	B_1	B_2
2	4	6	12
1 : 2		1 : 2	

Таблица 6.5 – Равномерный комплекс

A_1		A_2	
B_1	B_2	B_1	B_2
2	2	8	8
1 : 1		1 : 1	

Для равномерных и пропорциональных комплексов сумма частных дисперсий равна общей:

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

В некоторых случаях бывает легко организовать пропорциональный комплекс на основе имеющегося неравномерного. Рассмотрим, как это сделать на следующем примере.

Предположим, что имеется некоторое количество особей, которые по своему качеству отвечают требованиям градаций двухфакторного комплекса, но по количеству не отвечают требованиям пропорциональности (табл. 6.6).

Тут отношение частот фактора B по разным градациям фактора A неодинаково. Для A_1 отношение частот равно 1 : 3,3 : 4,3; для A_2 – 1 : 2,7 : 3,6. Но в этом неравномерном комплексе отношение частот по градациям каждого фактора в отдельности (что показано в правой части табл.6.6) близки к определенным целым числам:

$$A_1:A_2\approx1:3 \qquad B_1:B_2:B_3\approx1:3:4$$

Эти числа пропорциональности (k_{Ai}, k_{Bi}) можно использовать для построения пропорционального комплекса.

Таблица 6.6 – Исходные данные

	$A_1(1)$			$A_2(3)$			Сумма частот по градациям	Число пропорциональности
	B_1	B_2	B_3	B_1	B_2	B_3	3+10+13=26 (A_1)	$k_{A1}=\frac{26}{26}=1$
n_x	3	10	13	11	30	39		$k_{A2}=\frac{80}{26}=3{,}1\approx 3$
	$\frac{3}{3}\div\frac{10}{3}\div\frac{13}{3}=$ $=1\div 3{,}3\div 4{,}3$			$\frac{11}{11}\div\frac{30}{11}\div\frac{39}{11}=$ $=1\div 2{,}7\div 3{,}6$			11+30+39=80 (A_2)	$k_{B1}=\frac{14}{14}=1$
							3+11=14 (B_1)	$k_{B2}=\frac{40}{14}=2{,}9\approx 3$
							10+30=40 (B_2)	$k_{B3}=\frac{52}{14}=3{,}7\approx 4$
							13+39=52 (B3)	

Для этого в каждой градации по обоим факторам надо перемножить соответствующие числа пропорциональности, затем фактические частоты разделить на эти произведения и взять наименьшее из полученных частных (a_{min}). Эту величину надо умножить на произведения чисел пропорциональности. Таким образом, получают частоты пропорционального комплекса, который образован из имеющегося непропорционального с наименьшей выбраковкой особей. Эти действия показаны в следующей табл.6.7.

Таблица 6.7 - Организация пропорционального дисперсионного комплекса на основе имеющегося непропорционального

Градации 1-го фактора	$A_1(1)$			$A_2(3)$			
Градации 2-го фактора	$B_1(1)$	$B_2(3)$	$B_3(4)$	$B_1(1)$	$B_2(3)$	$B_3(4)$	
Фактические частоты n_x	3	10	13	11	30	39	n=106
Произведение чисел пропорциональности $\Psi=k_{Ai}\cdot k_{Bi}$	1·1=1	1·3=3	1·4=4	3·1=3	3·3=9	3·4=12	
$a=\frac{n_x}{\psi}$	3,0	3,3	3,3	3,7	3,3	3,3	a_{min}=3
$\psi\cdot a_{min}=n_x^*$	1·3=3	3·3=9	4·3=12	3·3=9	9·3=27	12·3=36	n^*=96
	1:3:4			1:3:4			

Однофакторный комплекс.

При изучении действия на результативный признак одного фактора, всегда присутствует только одна пропорция частот по градациям этого фактора, которую не с чем сравнивать. Поэтому для однофакторных комплексов отпадает требование пропорциональности или равномерности: однофакторные комплексы ортогональны при любом соотношении частот по градациям фактора. К однофакторным комплексам в полной мере относится требование рандомизации.

Двухфакторные и многофакторные комплексы.

В двухфакторных (многофакторных) комплексах необходимо обязательно иметь независимость изучаемых факторов и, желательно, пропорциональность в частотах. Как и все другие комплексы, двухфакторные должны быть рандомизированы.

Изучая действия более одного фактора, необходимо учитывать не только влияние каждого фактора в отдельности, но и их сочетаний, как было показано в примере выше, т.е. должны быть рассчитаны частные средние по фактору А, В, по сочетаниям факторов АВ и по их общему суммарному действию — по фактору х. По каждому ряду средних рассчитаны центральные отклонения, сумма квадратов которых дает дисперсию по каждому фактору.

7. РЕГРЕССИОННЫЙ АНАЛИЗ

Вначале рассмотрим несколько важных понятий регрессионного анализа.

Регрессией называется изменение функции (Y) при определенных изменениях одного или нескольких аргументов (х).

Функция – это признак, зависящий от других признаков - аргументов. Зависимость функции от аргументов может быть:

физиологической;

условно принятой в исследовании.

Примером физиологической зависимости может служить зависимость веса животного (функции) от его возраста (аргумента).

Если по длине определяется вес животного, считается что вес зависит от длины, если необходимо предусмотреть размеры животных разного веса, то принимается, что длина зависит от веса. Это пример условной зависимости. Вскрыть функцию – означает найти закономерности, по которым изменяется изучаемый признак в зависимости от изменения одного или нескольких признаков.

Если изменения функции исследуется в зависимости от одного признака, то регрессия называется простой:

$$Y = f(x) \Leftrightarrow x = f(y)$$

Если изучается зависимость изменения функции от изменения нескольких признаков, регрессия называется множественной.

$$Y = f(x_1, x_2, ..., x_n)$$

Если при одинаковом приращении аргумента, но при разных его значениях (малых, больших или средних) функция имеет неодинаковое приращение, причем среднее ее изменение не идет по прямой, то регрессия называется криволинейной (рис. 7.1).

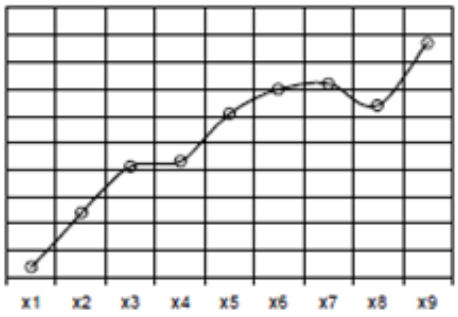


Рисунок 7.1 - Криволинейная регрессия

$$\Delta x_1 = \Delta x_2 = \ldots = \Delta x_n;\ \ \Delta y_1 \neq \Delta y_2 \neq \ldots \neq \Delta y_n;\ \ \frac{\Delta y_1}{\Delta x_1} \neq \frac{\Delta y_2}{\Delta x_2} \neq \ldots \neq \frac{\Delta y_n}{\Delta x_n} \neq \text{Const}$$

Если при любом значении (малом, среднем или большом) аргумента одинаковое изменение его приводит к одинаковому изменению значения функции, то регрессия называется прямолинейной (рис.7.2).

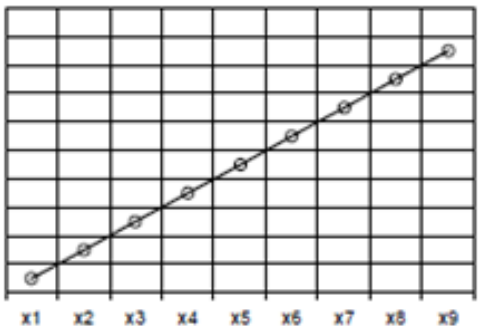


Рисунок 7.2 - Прямолинейная регрессия

$$\Delta x_1 = \Delta x_2 = \ldots = \Delta x_n;\ \ \Delta y_1 = \Delta y_2 = \ldots = \Delta y_n;\ \ \frac{\Delta y_1}{\Delta x_1} = \frac{\Delta y_2}{\Delta x_2} = \ldots = \frac{\Delta y_n}{\Delta x_n} = \text{Const}$$

На практике регрессию принято изображать в виде: регрессионного ряда (эмпирического или теоретического); линии регрессии (эмпирической или теоретической); коэффициентов регрессии, которые образуют уравнение регрессии, например:

$$Y = \sum_{i=0}^{n} a_i \cdot x^i = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_n \cdot x^n$$

Рассмотрим каждый из способов представления регрессии.

Эмпирический ряд регрессии – это двойной ряд цифр, который включает значения аргумента и соответствующие им значения функции, которые получены опытным путем. Пример эмпирического ряда регрессии дан в табл.7.1.

Таблица 7.1 - Пример эмпирического ряда регрессии

Возраст (x), годы	2	3	4	5	6	7	8	9	10
Живой вес (Y), кг	394	414	420	433	451	460	462	454	477

Составление эмпирического ряда регрессии. Для составления эмпирического ряда регрессии весь первичный материал разбивается на столько групп, сколько установлено градаций аргумента, и по каждой группе подсчитывается ∑V – общая сумма значений функции и n – число особей. Средняя получается простым делением первого числа на второе:

$$M_i = \frac{\sum V_i}{n}$$

При графическом изображении эмпирического ряда регрессии – аргумент, например возраст, откладывается по оси абсцисс, а функция, например вес, откладывается по оси ординат. В итоге получают эмпирическую линию регрессии (см.рис. 7.3).

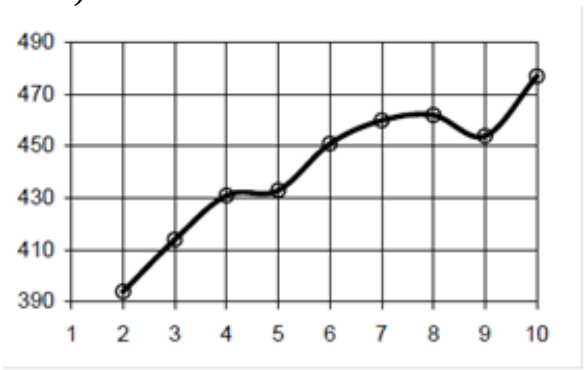


Рисунок 7.3 – Эмпирическая линия регрессии

Траектория эмпирической линии регрессии почти никогда не бывает плавной: в пределах одних интервалов аргумента, функция имеет повышенное, других – пониженное, а иногда и отрицательное приращение, что на графике дает ломанную кривую.

Ломанный характер эмпирической линии регрессии отражает обычную не выравненность общих условий развития признака-функции (веса) на различных участках изменения признака-аргумента (возраста). Если изучается регрессия веса по возрасту, то становится очевидным, что влияние всевозможных агентов на возрастные изменения веса не остается одинаковым на всем протяжении периода роста. В одном возрасте вся сумма влияний складывается в комплекс, более благоприятный для роста, в другом возрасте этот комплекс влияний не способствует достаточному приросту веса.

Таким образом, по виду эмпирической линии регрессии всегда можно установить на каких участках изменения аргумента признак-функция развивался в лучших, а на каких в худших условиях. Анализ эмпирической линии регрессии всегда дает практически ценную характеристику всех обстоятельств, которые связаны с зависимостью изучаемой функции от выбранного аргумента.

На практике, для нахождения основных форм зависимости функции от аргумента необходимо выяснить такое течение функции при равномерном изменении аргумента, которое соответствует усредненному, т.е. одинаковому влиянию всего комплекса условий, которые определяют развитие признака-функции.

Нахождение усредненного выровненного течения функции, в некоторой степени, подобно определению средней арифметической нескольких значений признака.

Средняя арифметическая получается путем сглаживания индивидуальных различий усредняемого признака, кроме этого она близко стоит ко всем индивидуальным значениям, так что сумма квадратов отклонений от их средней есть величина наименьшая. Эти же принципы положены и в основу нахождения усредненного течения функции.

Однако между усреднением течения функции и определением средней арифметической есть и существенные отличия: средняя арифметическая всегда имеет дело с одной переменной величиной (от особи к особи), а выровненное течение функции всегда имеет дело с двумя или несколькими переменными величинами, из которых одна (функция) величина зависит от других (аргументы) величин.

Процесс получения усредненного течения изменения функции при равномерном увеличении значения аргумента – называется выравниванием эмпирических рядов.

В результате выравнивания на основе эмпирической ломанной линии получается усредненная, плавная теоретическая линия регрессии, которая отражает основную закономерность зависимости функции от аргумента.

На практике выравнивание эмпирических рядов производится графически или аналитически.

При аналитическом методе выравнивания эмпирических рядов в результате составления уравнения регрессии первоначально вскрывается форма зависимости данной функции от выбранного аргумента. Подставляя в полученное уравнение регрессии последовательные значения аргумента, можно определить теоретический ряд значений функции, а нанося эти значения на график – получить теоретическую линию регрессии.

Для эмпирического ряда, который был рассмотрен в примере выше, уравнение регрессии имеет следующий вид:

$$y = a - b \cdot 10^{-c \cdot x} = 470 - 72 \cdot 10^{-0{,}1312 \cdot x}$$

где: y – теоретический вес особи; x – возраст в годах;

a=470 – максимальное значение веса, к которому асимптотически приближается данная функция по мере увеличения аргумента (возраста);

b=72 – сумма прироста от первого имеющегося значения возраста до его значения при остановке роста;

c=0,1312 – показатель темпа роста.

Теоретический и эмпирический ряды приведены в следующей табл.7.2.

Как известно из корреляционного анализа, в простейшем случае, при прямолинейной регрессии:

$$y=a \bullet x$$

зависимость функции от аргумента может быть выражена одним числом – коэффициентом регрессии, который показывает в каком направлении и насколько изменяется функция при увеличении значения аргумента на одну

единицу измерения. В природе существует множество явлений, которые обусловлены множеством причин. Поэтому существует очень много форм зависимости функций от различных аргументов. Исследование этих форм, выраженных математическими уравнениями, составляет основное содержание учения о регрессии признаков.

Таблица 7.2 – Теоретический и эмпирический ряды

Возраст (у), годы	2	3	4	5	6	7	8	9	10
эмпирический ряд									
Живой вес (х), кг	394	414	420	433	451	460	462	454	477
Теоретический ряд									
Живой вес (х), кг	398	417	431	441	449	454	458	461	464

Выравнивание эмпирических рядов регрессии имеет большое и разностороннее применение. Вскрывая усредненное течение функции, исследователь выявляет ту закономерность изучаемого явления, которая в эмпирическом ряду была вскрыта случайностями своего проявления. Эта закономерность, выраженная формулой или теоретическим рядом регрессии, помогает более точно, с меньшими ошибками дать описание внешних проявлений закономерности, что, в свою очередь, может помочь нахождению и внутренних факторов, управляющих данным явлением. В этом и заключается познавательное значение исследований регрессии различных признаков у эколого-биологических объектов.

Результаты этих исследований имеют также широкое применение и в практике. Каждый выровненный ряд дает возможность определить значение функции при любом значении аргумента (или нескольких аргументов). Это обстоятельство дает возможность использовать ряды и уравнения регрессии при определении значений таких признаков, непосредственное измерение которых в обычных условиях или невозможно, или затруднительно.

В практических работах использование уравнений и линий регрессии получило широкое распространение при определении без взвешивания, путем измерения, нормального живого веса животных и их убойного веса при жизни, веса сена в стогах, веса овощей в овощехранилищах, веса силосной массы в силосах, веса древесины в стволах и штабелях и др.

Широкое практическое применение во многих отраслях производства находит также специальная форма линий регрессии — номограмма.

7.1. Общие способы выравнивания эмпирических рядов

К общим способам выравнивания эмпирических рядов относятся:
графический способ;
способ скользящей средней (простой и взвешенной);
метод (способ) наименьших квадратов (МНК).
Далее рассмотрим применение каждого их этих способов в отдельности

7.1.1. Графический способ

Графический способ дает возможность с достаточным приближением получить теоретическую линию, а затем и теоретический ряд регрессии без каких-либо вычислений.

Наиболее простым оказывается применение графического способа к прямолинейной регрессии. В этих случаях на график наносится сначала эмпирическая линия регрессии, затем между крайними выступами ломаной эмпирической линии проводится прямая таким образом, чтобы сумма расстояний теоретической прямой от точек эмпирической линии была бы наименьшей.

При известном навыке это можно сделать от руки. При этом может помочь и натянутая нитка или прозрачная линейка с нанесенной прямой чертой. Натянутая нить располагается по среднему течению эмпирической линии, и после нахождения наилучшего положения нитки на графике отмечаются две крайние точки: для минимального и максимального значения аргумента. Теоретической линией регрессии будет прямая, соединяющая эти две точки.

По теоретической прямой можно определить числовые значения функции (ординаты), соответствующие определенным значениям аргумента (абсциссы).

Если регрессия не может считаться прямолинейной, то графическое выравнивание эмпирической кривом также может быть проведено, по для этого необходимо иметь представление об общих закономерностях изменения функции. Например, при изучении возрастных изменений живого веса сельскохозяйственных животных требуется учитывать, что живой вес, увеличиваясь с возрастом, постепенно приближается к некоторому максимальному значению, после чего прирост прекращается и значение его остается примерно на одном максимальном уровне.

7.1.2. Способ скользящей средней

Если форма функции неизвестна, то сгладить изломы эмпирической кривой можно, применив способ простой скользящей средней. Этот способ заключается в том, что для каждого значения аргумента берется средняя арифметическая из нескольких (соседних) значений функции.

Если скользящая средняя берется по трем значениям аргумента, то складываются значения функции для меньшего значения аргумента, для данного и для большего. Частное от деления этой суммы на 3 дает выровненное значение функции для данной величины аргумента.

Выравнивание эмпирического ряда методом простой скользящей средней показано в табл.7.3. Выровненная этим методом кривая дана на рисунке 7.4.

Выравнивание эмпирических рядов способом простой скользящей средней применяется, когда не требуется особой точности и когда имеется достаточно длинный ряд и можно пренебречь потерей двух значений функции, соответствующих крайним значениям аргумента.

Таблица 7.3 – Выравнивание эмпирического ряда по способу простой и взвешенной скользящей средней. (аргумент — содержание перевариваемого белка (%) в рационе телят до шестимесячного возраста, функция (y) — вес телят (кг) в возрасте шести месяцев)

Процент белка в рационе	Живой вес (y)	Сумма трех соседних (y)	Выровненные значения y_v по методу	
			«Простой»	«Взвешенной»
-	y_{+2}=0,5·(2·89,5+103-125)=78,5 – дополнительное значение			
-	y_{+1}=0,5·(2·103+120-147)=89,5 - дополнительное значение			
56	y_1=103	-	-	130,45
53	y_2=120	348*	116	117,25
50	y_3=125	392*	131	127,6
47	y_4=147	411	137	138,9
44	y_5=139	439	146	142,8
41	y_6=153	439	146	148,5
38	y_7=147	454	151	149,5
35	y_8=154	455	152	152,0
32	y_9=154	457	152	152,8
29	y_{10}=149	462	154	151,6
26	y_{11}=159	448	149	152,0
23	y_{12}=140	451	150	144,9
20	y_{13}=152	410	137	139,75
17	y_{14}=118	-		124,85
-	y_{n+1}=0,5·(2·118+152-159)=114,5- дополнительное значение			
-	y_{++2}=0,5·(2·114,5+118-140)=103,5 - дополнительное значение			

*$y_1+y_2+y_3$=103+120+125=348
$y_2+y_3+y_4$=120+125+147=392

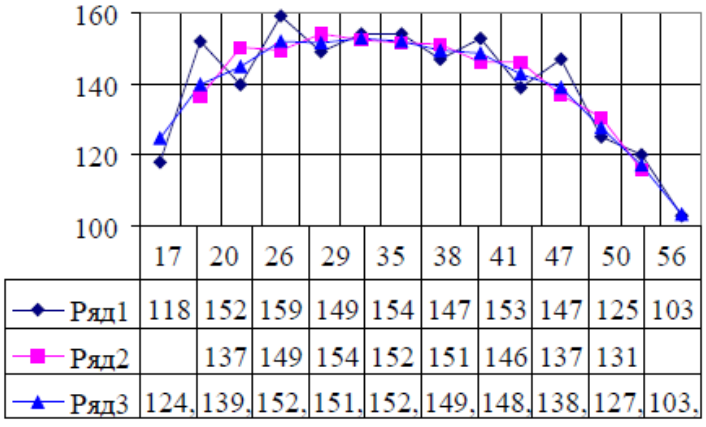


Рисунок 7.4 – Выравнивание эмпирического ряда способом простой и взвешенной средней

Более точные и не связанные с потерей крайних значений результаты получаются при использовании взвешенной скользящей средней. При этом

способе с обоих концов ряда добавляется по два значения — по два члена ряда. Определяются они следующим образом.

Первое (от конца) значение ряда (y_1) умножается на 2, к полученному произведению прибавляется второе значение (y_2), третье (y_3) пропускается, а из суммы вычитается четвертое (y_4). Полученное число делится на 2.Частное будет первым добавочным значением: y_{+1} (для начала ряда) или y_{n+1} (для конца ряда). Все эти действия можно выразить следующей формулой:

$$y_{+1} = \frac{2 \cdot y_1 + y_2 - y_4}{2}; \qquad y_{n+1} = \frac{2 \cdot y_n + y_{n-1} - y_{n-3}}{2}$$

Вторые добавочные значения с обоих концов ряда:

$$y_{n+2} = \frac{2 \cdot y_{+1} + y_1 - y_3}{2}; \qquad y_{n+2} = \frac{2 \cdot y_{n+1} + y_n - y_{n-2}}{2}$$

Для его расчета нужно использовать первое добавочное значение, а также первое и третье значения первоначального эмпирического ряда. Для рассматриваемого примера в двух верхних и в двух нижних строках табл.7.3 показано получение добавочных значений.

Для коротких рядов добавочные значения можно получать, пользуясь следующими формулами:

$$y_{+1} = \frac{4 \cdot y_1 + y_2 - 2 \cdot y_3}{3}; \qquad y_{+2} = \frac{4 \cdot y_{+1} + y_1 - 2 \cdot y_2}{3}$$

После установления добавочных значений приступают к выравниванию эмпирического ряда. Выровненные значения получаются путем вычисления взвешенной средней арифметической из пяти соседних эмпирических значений функции, взятых соответственно с весами 1; 2; 4; 2; 1.

Для того, чтобы получить, например, первое выровненное значение функции, нужно сумму второго добавочного, удвоенного первого добавочного, учетверенного первого эмпирического, удвоенного второго эмпирического и третьего эмпирического значений функции разделить на сумму весов (1+2 + 4 + 2+1 = 10).

Это можно выразить следующей формулой:

$$y_{v1} = \frac{y_{+2} + 2 \cdot y_{+1} + 4 \cdot y_1 + 2 \cdot y_2 + y_3}{10}$$

Для рассматриваемого эмпирического ряда первые выровненные значения функции будут равны:

$$y_{v1} = \frac{78{,}5 + 2 \cdot 89{,}5 + 4 \cdot 103 + 2 \cdot 120 + 125}{10} = 103{,}45$$

$$y_{v2} = \frac{89{,}5 + 2 \cdot 103 + 4 \cdot 120 + 2 \cdot 125 + 147}{10} = 117{,}25$$

В табл.7.3 приведен расчет всех выровненных значений функции для рассматриваемого примера. Ряд живого веса шестимесячных телят,

выровненный методами простой и взвешенной скользящей средней, показан на рисунке7.4.

7.1.3. Метод наименьших квадратов (МНК)

Наиболее распространенным общим аналитическим способом выравнивания эмпирических рядов регрессии является метод (способ) наименьших квадратов (МНК). Этот метод предоставляет наиболее универсальную возможность для выравнивания и определения вида аналитической функции (прямолинейной, обратной, параболической, гиперболической, степенной, логарифмической, экспоненциальной, периодической, простой, множественной и комбинации их) приближенно заменяющей табличные данные полученные экспериментальным путем.

МНК предназначен для выбора из совокупности назначенного типа кривых (прямолинейной, обратной, параболической и т.д.) такой кривой, для которой сумма квадратов отклонений эмпирических данных от выровненных (вычисленных по формуле данного типа кривой) является наименьшей, т.е.:

$$(y_e - y_p)^2 \Rightarrow \min$$

Рассмотрим на примере определение линейной зависимости МНК:

$$y = b \cdot x + a$$

Из условия минимума, для этой зависимости необходимо подобрать b и a так, чтобы отклонение экспериментальных точек от теоретической кривой было бы наименьшим, т.е.:

$$y_{Pi} = b \cdot x_i + a$$

$$\delta_i = y_{Эi} - y_{Pi}$$

$$y_{Pi} = y_{Эi} - \delta_i = b \cdot x_i + a$$

$$\delta_i = y_{Эi} - (b \cdot x_i + a)$$

Будем искать функцию с неизвестными b и a, у которой сумма σ_i была бы самой малой положительной величиной. При этом нужно учитывать, что возможны и отрицательные значения σ_i, поэтому будем искать экстремум минимума для параболы, т.е. возведем в квадрат обе части уравнения разности:

$$(\delta_i)^2 = (y_{Эi} - (b \cdot x_i + a))^2$$

Сумма квадратов отклонений экспериментальных данных от вычисленных составит:

$$S = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (y_{Эi} - (b \cdot x_i + a))^2$$

Составим систему уравнений поиска экстремума (минимума):

$$\begin{cases} \frac{\partial S}{\partial b} = 0 \\ \frac{\partial S}{\partial a} = 0 \end{cases}, \qquad \begin{cases} \sum_{i=1}^{n} 2 \cdot (y_{Эi} - (b \cdot x_i + a)) \cdot (-x_i) = 0 \\ \sum_{i=1}^{n} 2 \cdot (y_{Эi} - (b \cdot x_i + a)) \cdot (-1) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} x_i \cdot y_{Эi} - b \cdot \sum_{i=1}^{n} x_i^2 - a \cdot \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_{Эi} - b \cdot \sum_{i=1}^{n} x_i - a \cdot n = 0 \end{cases}$$

Чтобы решить эту систему уравнений сначала относительно b, умножим первое уравнение на n, а второе – на $\left(-\sum_{i=1}^{n} x_i \right)$:

$$n \cdot \sum_{i=1}^{n} x_i \cdot y_{Эi} - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_{Эi} - b \cdot \left(n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) = 0$$

Тогда:

$$b = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_{Эi} - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_{Эi}}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \qquad a = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_{Эi} - b \cdot \sum_{i=1}^{n} x_i \right)$$

В общем случае выравнивание эмпирических рядов регрессии МНК осуществляется по следующим этапам:

Определение общего вида уравнения регрессии. Производится на основе предварительного биологического анализа процессов, определяющих течение функции, или на основе рассмотрения эмпирической кривой.

Составление системы нормальных уравнений. Производится по следующим правилам, показанным на следующем примере. Пусть исходное уравнение $y = a + b \cdot x + c \cdot x^2$ изображается так, чтобы функция (y) была в правой части:

$$a + b \cdot x + c \cdot x^2 = y$$

Все члены исходного уравнения поочередно умножаются на величины, стоящие рядом с искомыми коэффициентами a, b, c, т.е. на 1, на x, на x^2:

$$(\times 1): \qquad a + b \cdot x + c \cdot x^2 = y$$

$$(\times x): \qquad a \cdot x + b \cdot x^2 + c \cdot x^3 = y \cdot x$$

$$(\times x^2): \qquad a \cdot x^2 + b \cdot x^3 + c \cdot x^4 = y \cdot x^2$$

У каждого слагаемого уравнения ставится знак суммирования. Логичным является вынести искомые коэффициенты a, b, c за этот знак, а также следует учитывать, что $\sum_{i=1}^{n} 1 = n$ – число пар аргумент-функция:

$$a \cdot n + b \cdot \sum x + c \cdot \sum x^2 = \sum y$$

$$a \cdot \sum x + b \cdot \sum x^2 + c \cdot \sum x^3 = \sum y \cdot x$$

$$a \cdot \sum x^2 + b \cdot \sum x^3 + c \cdot \sum x^4 = \sum y \cdot x^2$$

Полученные уравнения и есть система нормальных уравнений для данной исходной параболической функции. Эти правила применимы и для любой исходной формулы, например:

$$a + \frac{b}{x} = y$$

$$(\times 1): \quad a + b \cdot x + c \cdot x^2 = y \qquad a \cdot n + b \cdot \sum \frac{1}{x} = \sum y$$

$$\left(\times \frac{1}{x}\right): \quad \frac{a}{x} + \frac{b}{x^2} = \frac{y}{x} \qquad a \cdot \sum \frac{1}{x} + b \cdot \sum \frac{1}{x^2} = \sum \frac{y}{x}$$

Определение числового значения сумм, входящих в нормальные уравнения. Производится путем суммирования предварительно вычисленных рядов:

$$\sum x; \quad \sum x^2; \quad \sum x^3; \quad \sum x^4; \quad \sum y; \quad \sum y \cdot x; \quad \sum y \cdot x^2$$

Определение коэффициентов основного уравнения. Производится путем решения системы нормальных уравнений обычными алгебраическими приемами. В рассматриваемом примере (исходное равенство $y = a + b \cdot x + c \cdot x^2$) коэффициентами будут a, b, c.

Выравнивание эмпирических рядов регрессии МНК можно показать на следующих примерах.

7.2 Прямолинейные функции вида $y = b \cdot x + a$

Пример. В некоторых случаях можно выполнить определение возраста коров по числу колец на рогах. Связь между числами колец на рогах и возрастом возникает оттого, что каждый отел, происходящий обычно ежегодно, оставляет на рогах коровы кольцо, отражающее замедление роста рога в периоды глубокой стельности, когда главная масса питательных веществ тратится на питание плода. Поэтому если к среднему числу лет, прошедших до первого отела прибавить число колец, умноженное на средний межотельный период, то это и будет примерным возрастом коровы. Это можно выразить уравнением прямой:

$$y = b \cdot x + a ,$$

где а – число лет до первого отела;

х – число колец на рогах;

у – возраст коровы в годах;

b – средний межотельный период.

Зная исходное уравнение, можно составить систему нормальных уравнений. В данном случае она будет достаточно простой:

$$a \cdot n + b \cdot \sum x = \sum y$$

$$a \cdot \sum x + b \cdot \sum x^2 = \sum y \cdot x$$

или воспользоваться ранее полученными формулами для расчета а и b:

$$b = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$a = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i - b \cdot \sum_{i=1}^{n} x_i\right)$$

Таким образом, необходимо определить следующие четыре величины сумм:

$$\sum x; \quad \sum x^2; \quad \sum y; \quad \sum y \cdot x$$

Эти суммы приведены в табл.7.4.

Таблица 7.4 - Выравнивание эмпирического ряда регрессии возраста коров (у) по числу колец на рогах (х) МНК

Нахождение параметров y = b·x + a				Построение теоретического ряда		
x	y^*	x^2	yx	bx	y=b·x+a	$y-y^*$
11	13,3	121	146,3	10,955	13,4	+0,1
10	12,4	100	124,0	9,960	12,4	0,0
9	11,5	81	103,5	8,964	11,4	-0,1
8	10,5	64	84,0	7,968	10,4	-0,1
7	9,5	49	66,5	6,972	9,4	-0,1
6	8,3	36	49,8	5,975	8,4	+0,1
5	7,4	25	37,0	4,980	7,4	-0,1
4	6,5	16	26,0	3,984	6,4	-0,1
3	5,5	9	16,5	2,988	5,4	-0,1
2	4,4	4	8,8	1,992	4,4	0,0
1	3,4	1	3,4	0,996	3,4	0,0
Σ=66	Σ=92,7	Σ=506	Σ=665,8	-	-	-

$$b = \frac{11 \cdot 665{,}8 - 66 \cdot 92{,}7}{11 \cdot 506 - (66)^2} = +0{,}996 \qquad a = \frac{1}{11} \cdot (92{,}7 - 0{,}996 \cdot 66) = 2{,}45$$

Таким образом, теоретическое значение возраста по числу колец на рогах можно определить по формуле:

$$y = 0{,}996 \cdot x + 2{,}45$$

Теоретический ряд возраста по числу колец на рогах показан на рисунке 7.5.

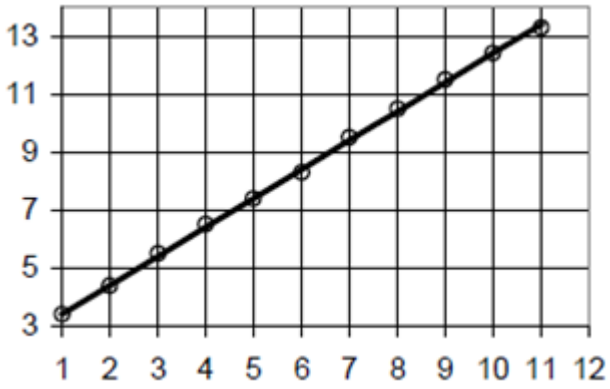


Рисунок 7.5 – Теоретический ряд возраста по количеству колец на рогах

Коэффициент корреляции для полученного уравнения рассчитываем по формуле:

$$r = \frac{\sum_{i=1}^{n}[(x_i - M_x) \cdot (y_i - M_y)]}{\sqrt{\sum_{i=1}^{n}(x_i - M_x)^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - M_y)^2}},$$

где Мх и Му – среднеарифметические соответственно для х и у:

$$M_x = \frac{\sum x}{n} = \frac{66}{11} = 6 \qquad M_y = \frac{\sum y}{n} = \frac{92{,}7}{11} = 8{,}43 \qquad r = 0{,}9997$$