МАТЕМАТИЧЕСКИЕ МЕТОДЫ В БИОЛОГИИ И ЭКОЛОГИИ

БАЗОВЫЕ ОПРЕДЕЛЕНИЯ

Статистическая совокупность – множество относительно однородных, но индивидуально различимых единиц, объединенных для группового изучения

Примеры: популяции животных, виды в экосистеме, стадо коров, растения на опытных делянках

Единица наблюдения (совокупности) – элемент статистической совокупности

Примеры: отдельная особь популяции, определенный вид в экосистеме, каждая корова в стаде, каждое отдельное растение на опытной делянке

Объем совокупности – число единиц наблюдения, входящих в статистическую совокупность

Примеры: число коров в стаде, число растений, отобранных для эксперимента, количество проб, в которых измерена концентрация загрязняющего вещества за определенный промежуток времени, количество животных, отловленных для изучения.

Признак – свойство, проявлением которого один объект отличается от другого (Лакин, 1990); характеристика объекта исследования

Примеры: цвет глаз, количество зерен в колосе, длина и вес особи, число видов в разных сообществах, процент жира в молоке определенной коровы, концентрация хлорофилла «а» в листьях растений.

В качестве синонимов используются такие термины, как *показатель*, *характеристика*, *величина*, употребление которых более приемлемо в отношении экологических объектов.

Генеральная совокупность и выборка.

Если исследование охватывает все единицы наблюдения статистической совокупности без единого исключения, то оно называется сплошным или полным (изучение всех особей биологической популяции, учет всех видов растений и животных в экосистеме). Если ограничиваются обследованием лишь некоторой части статистической совокупности, то исследование называется частичным или выборочным. В соответствии с этим в математической статистике принято делить статистическую совокупность на генеральную и выборочную.

Генеральная совокупность – это вся подлежащая изучению совокупность данных объектов. В пределе она рассматривается как состоящая из бесконечно большого количества отдельных единиц. Та часть объектов, которая подвергается исследованию, называется **выборочной совокупностью** или просто **выборкой**

Причины применения выборочного метода исследования:

1. Экономия времени, материальных и кадровых ресурсов при проведении исследования, поскольку изучается лишь часть генеральной совокупности.
2. Возможность изучать объекты, сплошное обследование которых практически невозможно или нецелесообразно.

ПРИЕМЫ ПЕРВИЧНОЙ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

Объектами исследований биологов и экологов могут быть системы различного уровня (клетка, орган, организм, популяция, биоценоз, экосистема) и разнообразные биологические (экологические) процессы и явления (размножение, питание, динамика численности популяций, сукцессия экосистем). Для изучения этих объектов необходимо получить, обработать и проанализировать соответствующие данные.
Данные – это исходная информация об объекте исследования, полученная путем наблюдения или эксперимента и представленная в форме, пригодной для постоянного хранения, передачи, обработки и анализа (например, набор конкретных чисел).
Зафиксированные сведения об изучаемом объекте представляют собой беспорядочную массу фактического материала, выраженного, как правило, в числовой, балльной, текстовой, знаковой формах. В современных условиях внедрения в научные исследования компьютерных технологий следующим обязательным этапом является ввод этих данных в одну из программ статистического анализа, к примеру в электронную таблицу MS EXCEL или пакет STATISTICA. Форма организации первичных данных в электронных таблицах при этом будет различаться в зависимости от цели статистической обработки.
Статистическая обработка первичных данных имеет и самостоятельное значение. Так, построение вариационных рядов и кривых может дать исследователю ценную информацию о законе распределения изучаемого признака или показателя, в дальнейшем это может помочь как в выборе корректных методов математической обработки, так и в определении факторов, вызывающих подобное вариационное распределение; вычисление средних значений и показателей вариации само по себе является важной характеристикой объекта исследований.

Статистические ряды

Математическая обработка собранных данных часто (но далеко не всегда!) начинается с построения так называемых **статистических рядов**, представляющих собой набор числовых значений признака, расположенных в определенном порядке.
Типы статистических рядов.
1. **Ранжированный ряд** – одинарный ряд, в котором значения признака располагаются в возрастающем (или убывающем) порядке.
Пример:
34342543345
23333444455 – ранжированный ряд.
2. **Вариационный ряд** (ряд распределения) – двойной ряд чисел, отражающий соотношение ранжированных значений признака с частотой их встречаемости в данной выборке.
Пример:
23333444455 – ранжированный ряд,
2345 – значение признака,
1442 – частота встречаемости.
Значение ряда: позволяет определить закономерность варьирования (закон распределения) изучаемого признака. В зависимости от того, в каком диапазоне и как варьирует признак – дискретно или непрерывно, – статистическая совокупность может распределяться в безынтервальный или интервальный вариационные ряды. Тип вариационного ряда можно определить по формуле (Лакин, 1990):

$$\mathbf{h} = \frac{\mathbf{x_{max} - x_{min}}}{\boldsymbol{k}},$$

где h – ширина классового интервала,
x_max; x_min – максимальное и минимальное значение выборки,

k – число классов, на которые следует разбить вариацию признака, рассчитывается по *формуле Стерджеса*:

$$k = 1 + \log_2(n),$$

где n – объем выборки.

Таким образом, если λ = 1 или λ ≈ 1, то строится безынтервальный ряд, если λ ≠ 1, то строится интервальный ряд. Частота встречаемости при этом может быть вычислена посредством функции ЧАСТОТА().

Пример

Рассмотрим данные о количестве птенцов в гнездах древесной ласточки Tachycineta bicolor (Рокицкий, 1973):

4 6 6 4 5 5 5 5 5 5 5 1 4 5 4 5 4 5 5 7 4 6 6 5 6 4 4 5 6 5 5 4 2 6 4 6 2 5 6 5 5 4

Данный признак является дискретным и λ ≈ 1, значит достаточно подсчитать встречаемость конкретных значений, не разбивая их на классовые интервалы. Искомый безынтервальный вариационный ряд будет выглядеть следующим образом:

Количество птенцов	Частота встречаемости
1	1
2	2
4	11
5	18
6	9
7	1

Интервальный вариационный ряд применяется, если изучаемый признак изменяется непрерывно (λ ≠ 1) или значения дискретного признака, варьирующего в широких пределах, имеют малую повторяемость.

В воде мелководного озера Неро (Ярославская область) в течение года были измерены концентрации общего фосфора (в мкг/л):

46	41	153	98	140	95	208	88	65	108	60	41	179
320	176	118	191	108	62	91	90	66	189	274	170	95
62	108	45	58	90	83	202	134	166	82	117	62	91
37	80	45	111	83	120	108	91	241	90	66	163	110
117	91	180	104	91	134	92	83					

Для построения интервального вариационного ряда сначала весь диапазон изменчивости концентраций общего фосфора разбивается на серию равных классовых интервалов, затем подсчитывается, сколько вариант попало в каждый интервал. В нашем примере ширина классового интервала λ = 41, число классовых интервалов k = 7, соответственно вариационный ряд имеет вид:

Классовые интервалы концентраций (мкг/л)	Частота встречаемости
37–78	14
78,1–119	28
119,1–160	5
160,1–201	8
201,1–242	3
242,1–283	1
283,1–324	1

3. **Временной ряд** (ряд динамики) – двойной ряд чисел, отражающий варьирование вариант изучаемого признака во времени (по годам, месяцам, дням, часам).

Пример: сезонные изменения биомассы фитопланктона в озере можно охарактеризовать следующим временным рядом

2	11	6	1	20	30	10	2 – биомасса фитопланктона (мг/л),
III	IV	V	VI	VII	VIII	IX	X – месяцы.

4. **Эмпирический ряд регрессии** – двойной ряд чисел, отражающий связь между значениями сопряженных признаков.

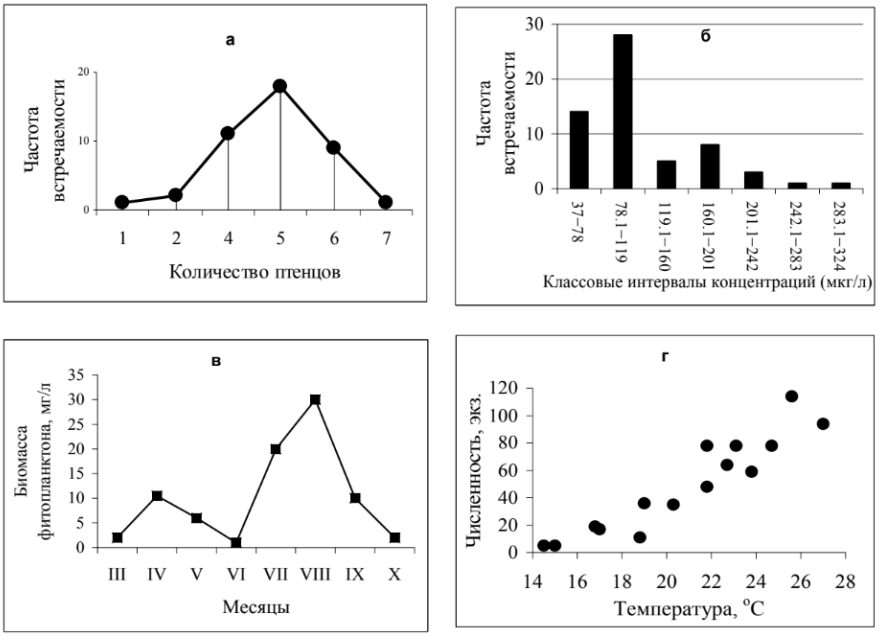
Пример: в районе биостанции «Улейма» студентами ЯрГУ были получены следующие данные о численности насекомых-опылителей на пробной площадке (X) и температуре воздуха в периоды учета насекомых (У):

X:	17	19	59	114	94	78	78	64	78	48	35	36	5	5	11
У:	17	16,8	23,8	25,6	27	24,7	21,8	22,7	23,1	21,8	20,3	19	15	14,5	18,8

Графический анализ

Визуализация, или наглядное представление, результатов исследований является важным этапом при первичной математической обработке данных.

Рис. 2.1. Графическое представление закономерностей статистических рядов: а – вариационная кривая распределения количества птенцов в гнездах древесной ласточки Tachycineta bicolor; б – гистограмма распределения концентраций общего фосфора; в – сезонная динамика биомассы фитопланктона в озере; г – точечная диаграмма, отражающая связь температуры воздуха и численности насекомых-опылителей на пробной площадке



Графическое представление закономерностей варьирования количественных признаков осуществляется с помощью **вариационных кривых** (полигон распределения частот) (рис. а) и **гистограмм распределения** (частот встречаемости значений признака) (рис. б).

Вариационные кривые строятся для *безынтервальных вариационных рядов* в осях: значения признака (абсцисса) – частота встречаемости значений признака (ордината). Данный график представляют собой ряд точек, соединенных прямыми линиями, при этом каждая точка отражает частоту встречаемости конкретного значения дискретного признака. Анализ вариационной кривой на рис. а обнаруживает характерную закономерность поведения количественного признака – число птенцов в гнездах древесной ласточки: высокие частоты встречаемости вариант наблюдаются в центре распределения, а низкие по периферии.

Весьма сходны с вариационными кривыми так называемые **гистограммы распределения частот** – столбчатые диаграммы, отражающие распределение частот встречаемости значений признака по отдельным классовым интервалам. Соответственно, в отличие от вариационной кривой на гистограмме распределения частот по оси абсцисс откладываются классовые интервалы. Подобные графики применяются для *интервальных вариационных рядов*. Возвращаясь к ранее описанному примеру, можно заключить, что закономерность варьирования концентраций общего фосфора значительно отличается от распределения количества птенцов в гнездах древесной ласточки: наблюдается смещение наиболее часто встречающихся концентраций фосфора в область меньших значений (рис. б). Табличный процессор MS EXCEL содержит процедуру автоматического построения из исходных данных одновременно вариационного ряда и гистограммы распределения частот этого ряда (**Анализ данных - Гистограмма**).

По данным *рядов динамики* строится **график** в осях: время (абсцисса) – значение признака (ордината) (рис. в). Графический анализ сезонной динамики биомассы фитопланктона выявляет наличие весеннего и позднелетнего пика в обилии микроводорослей. Спад в развитии приходится на ранее лето, в гидробиологии этот период именуется стадией «чистой воды», что часто связано либо с биогенным лимитированием, либо с выеданием фитопланктона зоопланктоном.

На основе *эмпирических рядов* регрессии строится **точечная диаграмма** (диаграмма рассеяния), отражающая связь между парой признаков (показателей) (рис. г). По оси абсцисс откладываются значения одного признака, по оси ординат – другого признака, сопряженного с первым. Таким образом, каждая точка на подобной диаграмме отражает значения пары признаков. Форма фигуры, создаваемой совокупностью точек на графике, является показателем связи двух признаков. Если между переменными существует сильная связь, то точки на графике образуют упорядоченную форму (например, близкую к прямой или кривой линии). Если переменные не связаны, то точки образуют «облако». Из рис. г видно, что точки образуют фигуру вытянутой формы, через которую в первом приближении можно провести прямую линию, при этом более высоким значениям температуры воздуха соответствуют более высокие численности насекомых-опылителей на пробной площадке. Это указывает на существование связи между двумя переменными.

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Статистическая обработка совокупности данных состоит в некоторых осредняющих вычислительных процедурах, погашающих сугубо индивидуальные особенности – отклонения от общей закономерности и подчеркивающих типичные (популяционные) свойства явления в целом. **Описательная статистика** – занимается характеристикой (описанием) картины случайного рассеяния по совокупности данных. В соответствии с законом распределения данных решаются вопросы выбора и вычислений надлежащих показателей. *К основным показателям описательной статистики относятся* среднее значение (среднее арифметическое, медиана, мода), усредненное значение, разброс (диапазон разброса данных), дисперсия, стандартное среднеквадратное отклонение (СКО), квартили, доверительный интервал.
Для проведения статистических расчетов в Microsoft Excel имеется надстройка **Пакет анализа**. Вычисления могут проводиться посредством команды **Данные - Анализ данных - Описательная статистика**, а также средствами *встроенных функций* MS Excel.

Лимиты (пределы вариации) – минимальное и максимальное значение признака в выборочной совокупности. Указывают границы варьирования признака.

MS EXCEL: МИН(), МАКС(), НАИМЕНЬШИЙ(), НАИБОЛЬШИЙ()

Размах вариации (интервал) – разность между максимальным и минимальным значением признака. Чем сильнее варьирует признак, тем больше показатели пределов и размаха вариации, и наоборот.

Объем выборки - количество измерений величин заданного признака.

MS Excel: СЧЁТ()

Сумма - сумма всех значений изучаемой переменной.

MS Excel: СУММ()

Меры центральной тенденции (средние величины)

Мера центральной тенденции – это число, характеризующее выборку по уровню выраженности измеренного признака. К мерам центральной тенденции относят:

- моду,
- медиану,
- среднее арифметическое.

Среднее (арифметическое) (M) - наиболее типичное (наиболее вероятное) значение в выборке, вокруг которого разбросаны все остальные значения признака, встречающиеся с меньшей вероятностью. Если значения интересующего нас признака у большинства объектов близки к их среднему и с равной вероятностью отклоняются от него в большую или меньшую сторону, лучшими характеристиками совокупности будут само среднее значение и стандартное отклонение. Напротив, когда значения признака распределены несимметрично относительно среднего, совокупность лучше описать с помощью медианы и процентилей.

MS Excel: СРЗНАЧ()

Медиана (Md) – значение признака, относительно которого ранжированный ряд делится на 2 равные части: в обе стороны от медианы располагается одинаковое число вариант. Медиана является важной характеристикой распределения случайной величины и, так же как математическое ожидание, может быть использована для центрирования распределения. Медиана определяется для широкого класса распределений (например, для всех непрерывных). Медиану и интерквартильный размах рекомендуется применять для описания распределения, *не являющегося нормальным* (а это большинство распределений медико-биологических параметров). Если имеется длинный хвост распределения, то медиана лучше, чем среднее значение, отражает «типичное» или «центральное» значение.

MS EXCEL: МЕДИАНА()

Мода (Mo) – значение признака, наиболее часто встречающееся в выборочной совокупности. Класс с наибольшей частотой называется модальным. На гистограмме распределения частот моде соответствует самый высокий столбец, на вариационной кривой – самая высокая точка.

MS EXCEL: МОДА(), МОДА.ОДН()

Начиная с MS EXCEL 2010 вместо функции МОДА() рекомендуется использовать функцию МОДА.ОДН() , которая является ее полным аналогом. Кроме того, в MS EXCEL 2010 появилась новая функция МОДА.НСК() , которая возвращает несколько наиболее часто повторяющихся значений (если количество их повторов совпадает). НСК – это сокращение от слова НеСКолько.

Медиана и *мода* менее чувствительны по сравнению со *средней арифметической* к крайним членам (наиболее низким и наиболее высоким значениям) выборочной совокупности, которые бывают для неё как раз менее характерными, и являются более устойчивыми характеристиками выборки.

Таким образом,

- для номинативных данных единственно подходящей мерой центральной тенденции является мода, т.е. та градация номинативной переменной, которая встречается наиболее часто.
- для порядковых и метрических переменных, распределение которых унимодальное и симметричное, мода, медиана и среднее совпадают. *При нормальном распределении Mo=Md=M*.
- Наиболее очевидной и часто используемой мерой центральной тенденции является среднее значение, но для его полноценного использования необходимо учитывать такие меры изменчивости признака, как среднее квадратичное отклонение (S) и ошибку средней (m).

Меры изменчивости признака (показатели вариации)

Меры центральной тенденции отражают уровень выраженности измеренного признака. Однако не менее важной характеристикой является выраженность индивидуальных различий испытуемых по измеренному признаку.

Меры изменчивости применяются для численного выражения величины межиндивидуальной вариации признака. К мерам изменчивости признака относят:

- дисперсию,
- стандартное отклонение,
- асимметрию,
- эксцесс.

Дисперсия (σ^2 , S^2) – мера рассеянности случайной величины (переменной), которая равна отношению суммы квадратов отклонений отдельных значений признака от средней арифметической к объему выборки за вычетом единицы

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

Дисперсия измеряется в квадратах единицы измерения. В самостоятельном виде (как, например, средняя арифметическая) дисперсия используется редко. Это скорее вспомогательный и промежуточный показатель, который применяют в других методах статистического анализа.

MS EXCEL: ДИСП(), ДИСП.В()

Свойства дисперсии:

1 Если значения измеренного признака не отличаются друг от друга (равны между собой), дисперсия равна нулю.

2 Прибавление одного и того же числа к каждому значению переменной не меняет дисперсию. Прибавление константы к каждому значению переменной сдвигает график распределения этой переменной на эту константу (меняется среднее), но изменчивость (дисперсия) при этом остается неизменной.

3 Умножение каждого значения переменной на константу с изменяет дисперсию в c^2 раз.

4 При объединении двух выборок с одинаковой дисперсией, но с разными средними значениями дисперсия увеличивается.

Среднее квадратическое (стандартное) отклонение (σ, S) – корень квадратный из дисперсии. Если стандартное отклонение рассчитывается по выборочным данным, то используется обозначение S, если на основе генеральной совокупности, то символ σ (читается как «сигма»). В итоге стандартное отклонение является в ряде случаев более удобной характеристикой вариации признаков, поскольку измеряется в тех же единицах, что и исходные данные.

MS EXCEL: СТАНДОТКЛОН(), СТАНДОТКЛОН.В()

Таким образом, дисперсия и стандартное отклонение являются мерой варьирования числовых значений признака вокруг их средней арифметической и одновременно отражают внутреннюю изменчивость значений признака, зависящую от разностей между отдельными значениями признака. *Однако эти показатели затруднительно использовать при решении ряда задач сравнения признаков по степени варьирования*. Поэтому в биологии и экологии широкое распространение получила также относительная количественная характеристика вариации.

Коэффициент вариации (C_v) – отношение стандартного отклонения к средней арифметической величине, **выраженное в процентах**:

$$C_v = \frac{S}{\bar{X}} \cdot 100\% .$$

Варьирование считается слабым при $C_v \le 10\%$, средним - при C_v от 11 до 25 % и сильным - при $C_v > 25$ %.
Дисперсия и *стандартное отклонение* применимы для сравнительной оценки признаков, выраженных в одних и тех же единицах измерения. *Коэффициент вариации* позволяет сравнивать вариацию признаков, выраженных разными единицами измерения. *Коэффициент вариации* позволяет также сравнивать вариацию признаков, выраженных в одних и тех же единицах измерения, но резко различающихся по величине среднего значения.
В статистике принято считать, что совокупность данных является однородной, если коэффициент вариации менее 33%, неоднородной – если более 33%. Эта информация может быть полезна для предварительного описания данных и определения возможностей проведения дальнейшего анализа.

Верхний (нижний) квартиль - такое значение случайной, больше (меньше) которого 25% значений выборки.

MS EXCEL: КВАРТИЛЬ(), КВАРТИЛЬ.ВКЛ(), КВАРТИЛЬ.ИСКЛ()

КВАРТИЛЬ.ВКЛ(массив;кварт), где
Массив Массив или диапазон ячеек с числовыми значениями, для которых определяется значение квартиля.
Часть Значение, которое требуется вернуть, в частности

Возможное значение аргумента Часть	КВАРТИЛЬ.ВКЛ возвращает
0	Минимальное значение
1	Первый (верхний) квартиль (25-ю персентиль)
2	Значение медианы (50-ю персентиль)
3	Третий (нижний) квартиль (75-ю персентиль)
4	Максимальное значение

Стандартная ошибка среднего значения

Стандартная ошибка среднего является интервальной оценкой («от – до» или ±) генерального среднего значения и рассчитывается на основе известного выборочного среднего.

Стандартная ошибка (средняя, среднеквадратическая, статистическая ошибка; ошибка репрезентативности, ошибка выборочности) – это средняя величина отклонения выборочной характеристики от её генерального параметра. Наиболее часто в биологии и экологии используется понятие «стандартная ошибка среднего значения».

Стандартная ошибка по своей природе является не ошибкой измерения, а статистической ошибкой, неизбежно возникающей при отборе выборок из генеральной совокупности и, соответственно, связанной с перенесением результатов, полученных при изучении выборки, на всю генеральную совокупность. При этом очевидно, что ошибки измерения могут увеличивать стандартную ошибку. Также следует понимать, что определять величину ошибок репрезентативности требуется только для выборочных характеристик, генеральные параметры не имеют стандартных ошибок.

Расчет стандартной ошибки фактически совпадает с вычислением стандартного отклонения, произведенного для выборки. Поэтому стандартная ошибка не что иное, как стандартное отклонение множества случайных выборочных средних от истинной генеральной средней. На практике обычно нет возможности делать несколько выборок и вычислять несколько выборочных средних, чтобы по ним проводить расчеты. Статистическая теория показывает, что стандартная ошибка среднего значения в n раз меньше, чем стандартное отклонение. Поэтому ошибку можно рассчитать для единичной отдельной выборки (на основе всего одного
выборочного среднего значения) по формуле:

$$S_{\bar{X}} = \frac{S_X}{\sqrt{n}},$$

где Sx – стандартное отклонение, n – объем выборки.

Полезно знать, что приведенная формула всегда даёт значения стандартной ошибки, несколько завышенные по сравнению с действительными, поскольку в расчетах используется выборочное стандартное отклонение, а не истинное генеральное. Данная неточность при расчетах стандартной ошибки считается допустимой, поскольку в статистике из альтернативы – преувеличение или преуменьшение ошибки – именно первое является менее опасным.

Формула стандартной ошибки показывает, что величина ошибки тем больше, чем больше варьирование признака (Sx) и чем меньше выборка (n). Таким образом, стандартная ошибка указывает на точность, с какой выборочный показатель характеризует генеральный параметр. Чем меньше ошибка, тем ближе выборочная характеристика к величине генерального параметра, и, наоборот, чем больше ошибка, тем менее точно выборочная характеристика определяет генеральный параметр, и значит пользоваться подобными данными необходимо с особой осторожностью.

Доверительный интервал для среднего значения

Стандартная ошибка характеризует лишь средние пределы варьирования выборочных средних около истинного генерального среднего значения.

Доверительный интервал – границы, в которых с заданной вероятностью (степенью достоверности) находится изучаемый генеральный параметр. В экологии и биологии наиболее часто используется доверительный интервал для среднего значения.

В случае с доверительными интервалами важнейшее значение имеет так называемое «соглашение о 95%-й вероятности». В соответствии с ним совокупности, состоящей из 95%
особей (объектов), мы доверяем так же, как и 100%-й. Другими словами, данная вероятность принята как наименьшая, которой можно доверять как 100%-й при принятии того или иного решения, связанного с математической обработкой данных. Поэтому подобная вероятность получила обозначение «доверительная вероятность» – вероятность, признанная достаточной для уверенного суждения о генеральных параметрах на основании известных выборочных характеристик. Применительно к

доверительному интервалу это вероятность того, что генеральный параметр (среднее значение) действительно окажется внутри доверительного интервала. Если вероятность 0,95 является наименьшей в рейтинге доверия, то, значит, существуют и другие доверительные вероятности. Действительно, если решение, которое нужно принять при математической обработке данных, является очень ответственным, то его стараются принимать с ещё большей вероятностью, к примеру 0,99 или 0,999, чтобы свести возможные ошибки практически к нулю. Все эти три вероятности относятся к доверительным, и именно они используются при построении доверительных интервалов.

Уровень значимости (p или α) – это вероятность того, что генеральный параметр (среднее значение) при заданной доверительной вероятности (P=0,95, P=0,99, P=0,999) окажется за границами доверительного интервала.

Отсюда более конкретное определение применительно к доверительному интервалу – это вероятность того, что генеральный параметр (среднее значение) при заданной доверительной вероятности (P=0,95, P=0,99, P=0,999) окажется за границами доверительного интервала.

В статистике приняты 3 уровня значимости, соответствующие доверительным вероятностям:

P = 0,95 p = 0,05 = 1-0,95
P = 0,99 => p = 0,01 = 1-0,99
P = 0,999 p = 0,001 = 1-0,999

Отсюда следует, что

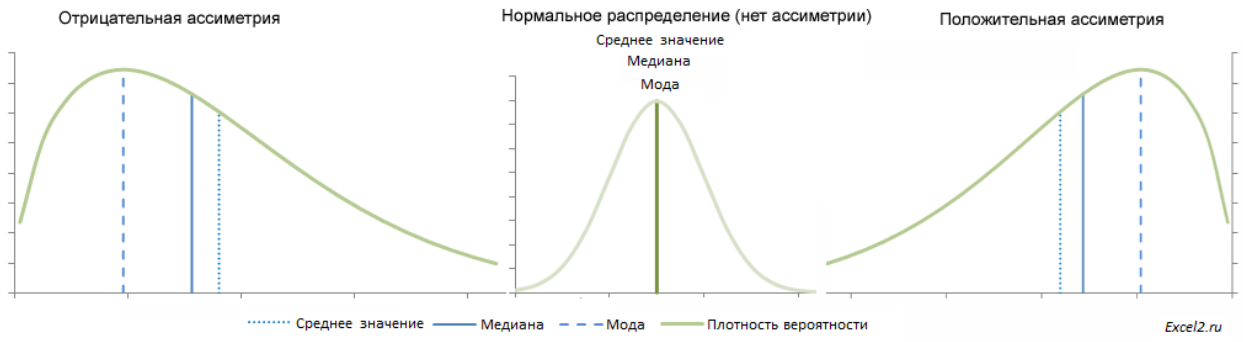
при p = 0,05 риск ошибиться составляет 1 раз на 20 случаев (5%),
при p = 0,01 – 1 раз на 100 случаев (1%),
при p = 0,001 – 1 раз на 1000 случаев (0.1%).

Таким образом, чем меньше уровень значимости и, соответственно, выше доверительная вероятность, тем меньше риск ошибки.

Важно понимать, что чем выше доверительная вероятность, тем шире будет доверительный интервал и тем менее чёткой становится оценка генерального параметра. В большинстве экологических и биологических исследований достаточно надежной считается 95%-я доверительная вероятность (или 5% уровень значимости), которые и используются наиболее часто.

MS EXCEL: ДОВЕРИТ.СТЬЮДЕНТ()

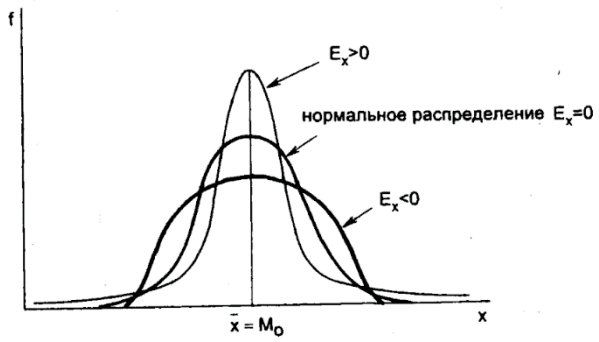
Асимметрия и и эксцесс распределения



Асимметричность (коэффициент асимметрии) характеризует степень несимметричности распределения (плотности распределения) относительно его среднего.

При нормальном распределении *мода* и медиана совпадают. При асимметрии *мода* отклоняется от *медианы* вправо либо влево. Если *мода* отклоняется влево от *медианы*, а правая ветвь кривой длиннее левой (т. е. является более пологой), то говорят о положительной (правосторонней) асимметрии, при этом коэффициент асимметрии As > 0. Если *мода* отклоняется вправо от *медианы*, а левая ветвь кривой длиннее правой, то говорят об отрицательной (левосторонней) асимметрии, при этом коэффициент асимметрии As < 0.

MS EXCEL: СКОС()



Эксцесс (Ex) – мера остроты пика распределения случайной величины. Эксцесс характеризует степень концентрации вариант вокруг среднего значения и является своеобразной мерой крутизны угла наклона кривой распределения признака. Если вершина эмпирической кривой оказывается сильно поднятой относительно вершины нормальной кривой, то говорят о положительном эксцессе распределения, при этом коэффициент эксцесса Ex > 0. Если вершина эмпирической кривой оказывается ниже вершины нормальной кривой, то говорят об отрицательном эксцессе распределения, при этом коэффициент эксцесса Ex < 0. Отрицательным пределом величины эксцесса является число -2, положительного предела нет.

MS EXCEL: ЭКСЦЕСС()

Причины возникновения асимметричных эмпирических распределений:

1. «Механическая» причина – асимметричность распределения связана с неправильной группировкой значений признака по классовым интервалам или с неправильным расчетом ширины классового интервала. В результате с одной стороны кривой частота встречаемости значений признака может оказаться больше, чем с другой. Подобные асимметричные распределения называются ложными.
2. Модифицирующие условия внешней среды – действие экстремальных или специфических факторов приводит к отклонению значений биологических признаков организма от «нормы», и при изучении данных признаков распределение их оказывается асимметричным.
3. Неоднородность выборки – объединение 2 разнородных совокупностей, каждая из которых имеет нормальное распределение, но в сумме они образуют асимметричное распределение (как правило, бимодальное или двухвершинное, в общем случае – полимодальное).

ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

Биологические и экологические явления (события) случайны, точно не предсказуемы. Начиная биологический эксперимент или приступая к наблюдению, невозможно точно сказать, каков будет результат – уровень численности животных в данном районе, выживаемость подопытных особей, артериальное давление через час после введения препарата. Поэтому биологам и экологам часто приходится сталкиваться с вероятностными (стохастическими) суждениями. Так, для гидробиолога, изучающего чистое олиготрофное озеро, ясно, что вероятность обнаружить массовое развитие водоросли Planktothrix agardhii (индикатор высокой степени сапробности) крайне мала. Эксперимент по проверке токсичности определенного вещества может показать, что в контрольном варианте выжило на 10% особей больше, чем в опытном (с добавкой вещества). Зависела ли эта разница в выживаемости особей от действия вещества или могла определяться другими факторами (например, изначальной разницей физиологического состояния особей в группах)? Экспериментатор может сказать следующее: «Очень вероятно, что именно тестируемое вещество определило большую смертность особей в опытной группе по сравнению с контрольной». Его более скептически настроенный коллега может заявить: «Небольшая разница, всего лишь в 10%, могла быть следствием действия случайных (неконтролируемых в эксперименте) причин, поэтому маловероятно, что вещество является токсичным».

Однако любому биологу и экологу ясно, что случайность изучаемых ими явлений относительна, несмотря на то, что точный прогноз невозможен, приблизительный результат можно предсказать. Каким образом можно дать такого рода прогноз?

Рассмотрим пример. Зоолог, изучающий популяцию какого-либо вида животного, задался целью дать приблизительный прогноз появления особей в популяции с некой мутацией (например, связанной с окраской). Чтобы рассчитать вероятность, ему потребуются предварительные исследования и данные о том, насколько часто в популяции рождаются особи с данной мутацией. Так, если исследователь обнаружит, что за ряд предшествующих лет из 10 000 родившихся особей 100 имели данную мутацию, то он сможет рассчитать вероятность рождения
мутантной особи в данной популяции:

$$p=100/10000=0,01$$

Другими словами, в среднем из 100 родившихся особей одна может быть мутантной. При наличии подобных данных можно решить и обратную задачу – найти вероятность появления в популяции особи без данной мутации:

$$p=9900/10000=0,99$$

Из этого примера вытекают 2 важных вывода.

1) сумма вероятностей противоположных событий (0.01+0.99) всегда равна единице.

2) приблизительный (вероятностный) прогноз можно дать, ориентируясь на повторяемость однотипных событий, на частоту встречаемости значений признака. Зная частоту, с которой данное значение признака встречается в популяции относительно общего количества всех встреченных значений признака (объем выборки), можно установить статистическую вероятность появления данного значения признака:

$$p=f/n$$

где f – частота встречаемости, n – объем выборочной совокупности.

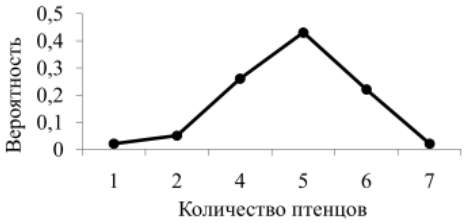
Статистическую вероятность события принято называть **относительной частотой**. Установлено, что относительная частота полностью не совпадает с «классической» вероятностью, однако приближается к ней по мере значительного увеличения числа наблюдений, т. е. объема выборки. Таким образом, зная ряд распределения частоты встречаемости значений признака, можно легко перейти к построению распределения вероятностей.

Вернемся к примеру о количестве птенцов в гнездах древесной ласточки

Количество птенцов	Частота встречаемости	Вероятность
1	1	1/42≈0,02
2	2	2/42≈0,05

4	11	11/42≈0,26
5	18	18/42≈0,43
6	9	9/42≈0,22
7	1	1/42≈0,02
Итого	42	1

Кроме того, закономерность, отмеченную в распределении вероятностей, можно выразить не только в табличной форме (ряд распределения), но и графически, построив кривую распределения вероятностей появления в выводке того или иного количества птенцов:



Наконец, закономерность распределения вероятностей можно описать с помощью математической формулы. Функция, связывающая значения случайного признака с их вероятностями, называется **законом распределения** признака. Каждый признак (показатель) распределяется по своему закону, имеет специфическую закономерность распределения (повторяемости) отдельных значений. Поэтому закон распределения образно можно сравнить с «паспортом» признака. В зависимости от типа переменной выделяют дискретные и непрерывные законы распределения. Описанное распределение относится к дискретным и, вероятнее всего, близко к так называемому биномиальному распределению.

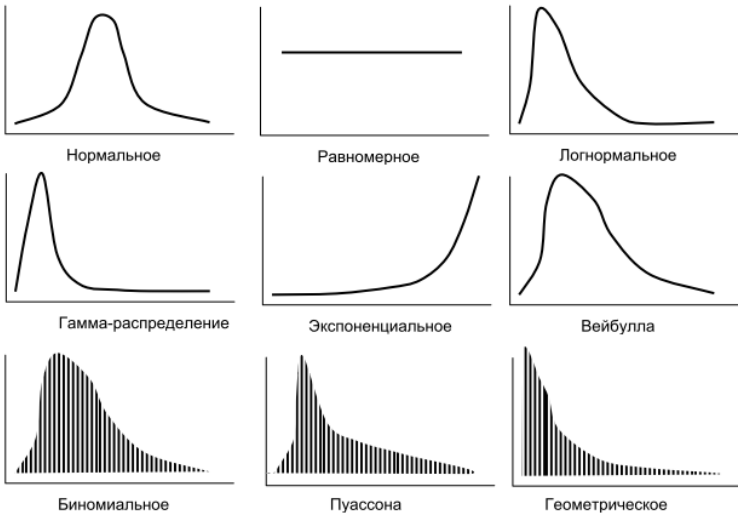


Рис. 3.2. Некоторые типы теоретических распределений случайной величины: *непрерывные* – нормальное, логнормальное, гамма-распределение, экспоненциальное, распределение Вейбулла; *дискретные* – биномиальное, распределение Пуассона, геометрическое, равномерное (по: Шитиков и др. 2003)

К настоящему моменту известны десятки теоретических распределений (их можно построить на основе известных математических формул), к которым исследователи могут «подгонять» полученные на основе выборок эмпирические распределения, устанавливая с определенной вероятностью, по какому закону распределяются изучаемые признаки. Из всего многообразия законов распределения кратко остановимся на наиболее значимых в практике биологических и экологических исследований – нормальном и биномиальном.

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Статистическая гипотеза

Под **статистической гипотезой** понимают формальное предположение о том, что сходство или различие некоторых параметрических или функциональных характеристик случайно или, наоборот, неслучайно

В биологических и экологических исследованиях анализ отдельных выборок редко является конечной целью. Очень часто приходится сравнивать эти выборки между собой. При сравнении двух выборок всегда возникает вопрос, достоверны ли наблюдаемые отличия между выборками или они обусловлены лишь какими-то случайными причинами? Другими словами, можно ли данное различие считать закономерным, характерным для всей генеральной совокупности и рассматривать его как результат реально действующих в системе факторов или же оно случайно и является следствием недостаточного количества данных и в следующих опытах (наблюдениях) может не проявиться?

Достоверность (статистическая значимость) – это свойство выборочной разности (различие средних, дисперсий 2-х выборок) правильно с заданной вероятностью отражать генеральную разность (различие генеральных средних и дисперсий). Выборочная разность может быть достоверна (статистически значима) или недостоверна (случайна, статистически незначима).

Выборочная разность достоверна – это означает, что если в выборочном исследовании зафиксировано различие выборочных характеристик (средних значений, дисперсий), то точно
такое же различие наблюдается между соответствующими генеральными параметрами в генеральных совокупностях, из которых извлечены выборки.

Если получена **недостоверная выборочная разность**, это значит не получено никакого определенного ответа о разности между соответствующими генеральными параметрами в генеральных совокупностях, из которых извлечены выборки. Другими словами, ничего нельзя заключить с заданной вероятностью о генеральной разности – ни что она есть, ни что её нет, т.е. *разница остаётся статистически недоказанной*.

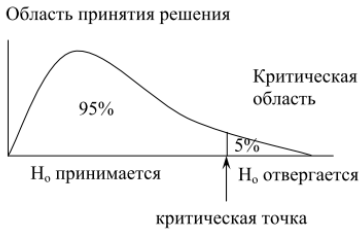
Распространенная ошибка среди исследователей – это неправильная интерпретация недостоверности различий выборочных характеристик: наличие между выборками недостоверной разности не свидетельствует об отсутствии разности между соответствующими генеральными параметрами, *фактически отсутствие различий в генеральных совокупностях доказать невозможно*.

Для установления того, достоверна или недостоверна выборочная разность, исследователь вначале должен сформулировать 2 противоположные статистические гипотезы:

1. **Нулевая гипотеза** (Но) – различия между выборочными характеристиками случайны, недостоверны.
2. **Альтернативная гипотеза** (На) – различия между выборочными характеристиками достоверны, т. е. реально наблюдаются между генеральными параметрами в генеральных совокупностях, из которых извлечены выборки.

Для отклонения или принятия той или иной гипотезы применяются так называемые **критерии достоверности** – специально разработанные статистические показатели с известными функциями распределения, позволяющие с заданной доверительной вероятностью проверять истинность нулевой или альтернативной гипотез.

Уровень значимость (p) – это вероятность ошибочного отклонения нулевой гипотезы при принятии решения о существовании различий, вероятность того, что результаты не представляют популяцию.



В частности, пусть в качестве критического принят уровень значимости P= 0,05, которому соответствует достоверность различий на уровне 95%. Если P≤0,05 – различия считаются достоверными. Если P>0,05 – различия недостоверны.

Статистические критерии

Статистический критерий – инструмент определения уровня статистической значимости. Статистические критерии обозначают также метод расчета определенного числа и само это число.

Все критерии используются с одной главной целью: определить уровень значимости анализируемых с их помощью данных (т.е. вероятность того, что эти данные отражают истинный эффект, правильно представляют популяцию, из которой сформирована выборка).

Все критерии различаются по мощности. **Мощность критерия** – это его способность выявлять различия или отклонять нулевую гипотезу, если она неверна.

Критерии можно разделить на две группы:

- параметрические
- непараметрические.

1. **Параметрические критерии** – рассчитываются на основе параметров выборочной совокупности (на основе среднего значения, дисперсии, стандартной ошибки и т. д.).
T-критерий Стьюдента (t-test) – основан на выборочном среднем значении. Если возникает задача сравнить две выборки по средним значениям, то фактическое значение критерия Стьюдента рассчитывается в классическом варианте как отношение разности 2-х выборочных средних к стандартной ошибке этой разности:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}.$$

При этом очевидно, что чем больше разность между средними значениями 2-х выборок и чем меньше стандартная ошибка этой разности, тем больше вероятность того, что 2 выборки достоверно различаются между собой по средней тенденции. Как отмечалось выше, для каждого подобного фактического значения критерия можно рассчитать фактический p-уровень значимости, сравнить его с критическим уровнем (0,05) и определить достоверность различий. Подобная схема принятия решения сохраняется для всех других критериев достоверности.

MS Excel: Парный двухвыборочный t-тест для средних, Двухвыборочный t-тест с различными дисперсиями, Двухвыборочный t-тест с одинаковыми дисперсиями (Анализ данных)

F-критерий Фишера (F-test) – основан на выборочной дисперсии. Фактическое значение F-критерия рассчитывается как отношение большей выборочной дисперсии к меньшей:

$$F = \frac{S_1^2}{S_2^2}.$$

MS Excel: Двухвыборочный F-тест для дисперсии (Анализ данных)

2. **Непараметрические критерии** – рассчитываются на основе частоты встречаемости или рангов.
Частотные критерии:

Критерий χ² («хи квадрат»), или критерий согласия Пирсона, (Chi-Square test) – представляет собой сумму квадратов отклонений эмпирических частот (f) от вычисленных теоретических частот распределения (f'), отнесенную к теоретическим частотам:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f - f')^2}{f'}.$$

В зависимости от типа выборки:

1. Критерии для независимых выборок (t-критерий Стьюдента, критерий Манна – Уитни, критерий серий Вальда – Вольфовица).
2. Критерии для зависимых выборок (парный t-критерий Стьюдента, критерий Вилкоксона, критерий знаков).

Независимые выборки – сравниваемые выборки, отдельные значения в которых никак не связаны между собой. Объем обеих выборок может различаться

Пример: сравнение роста или веса животных из 2-х популяций, сравнение концентраций вещества за разные годы.

Зависимые выборки – сравниваемые выборки, отдельные значения которых попарно связаны между собой. Объем обеих выборок всегда должен быть равным . Преимущество подобных выборок в том, что при сравнении различия внутри выборок становятся меньшими, чем между выборками, это повышает вероятность установления достоверности выборочной разности.

Пример: сравнение силы левой и правой руки у группы испытуемых, сравнение физиологических показателей у одних и тех же животных до и после проведения опыта. Могут быть и более слабые варианты зависимости. Например, мужья — одна выборка, их жены — другая выборка (при исследовании, например, их предпочтений), или дети 5—7 лет — одна выборка, а их братья- или сестры — другая выборка.

Рекомендации к выбору критерия различий:

- прежде всего, следует определить, является ли выборка связной
- (зависимой) или несвязной (независимой);
- следует определить однородность – неоднородность выборки;
- оценить объем выборки и, зная ограничения каждого критерия по объему, выбрать соответствующий критерий;
- если в распоряжении имеется несколько критериев, то следует выбирать те из них, которые наиболее полно используют информацию, содержащуюся в экспериментальных данных;
- при малом объеме выборки следует увеличивать величину уровня значимости (не менее 1%), так как небольшая выборка и низкий уровень значимости приводят к увеличению принятия ошибочных решений.

Алгоритм оформления выводов

(вопросы, на которые необходимо сформулировать ответ):

1 Что анализировалось (какие испытуемые, параметры какой методики).
2 Посредством чего проводился анализ (какие критерии и методы анализа использовались).
3 Какова достоверность полученных результатов (на каком уровне с указанием либо его точного значения (p=0,03), либо той зоны, в которую это значение попадает ($p \le 0,05$)).
4 Интерпретация (что это означает в контексте данного исследования и какой вывод из этого следует сделать).

Методы сравнения средних значений

Сравнение средних значений различных выборок относится к наиболее часто применяемым методам статистического анализа. При этом всегда должен быть выяснен вопрос, *можно ли объяснить имеющееся различие средних значений статистическими колебаниями или нет*. Если нет, говорят **о статистически значимом различии** между сравниваемыми группами.

При сравнении средних значений выборок предполагается, что обе выборки подчиняются нормальному распределению. Если это не так, то вычисляются медианы и для сравнения выборок используется непараметрические тесты.

Сравнение двух средних может проводиться по

- критерию Стьюдента,
- критерию Фишера

При сравнении средних двух независимых выборок по критерию Стьюдента предварительно необходимо проверить, не различаются ли дисперсии (**Анализ данных - Двухвыборочный F-тест для дисперсии (критерий Фишера)**). В случае равенства дисперсий для вычислений применяется **Двухвыборочный t-тест с одинаковыми дисперсиями**, а в случае неравенства – **Двухвыборочный t-тест с различными дисперсиями**

При обработке двух зависимых выборок сравнение средних следует проводить, используя **парный t-критерий** (**Анализ данных - Парный двухвыборочный t-тест для средних**). В отличие от обычного критерия Стьюдента, парный t-критерий работает не с выборочными средними и выборочными дисперсиями, а со средней разностью и дисперсией разности между отдельными наблюдениями.

ДИСПЕРСИОННЫЙ АНАЛИЗ

Цель дисперсионного анализа - исследование наличия или отсутствия существенного влияния какого-либо качественного или количественного фактора на изменения исследуемого результативного признака. Для этого фактор, предположительно имеющий или не имеющий существенного влияния, разделяют на классы градации (говоря иначе, группы) и выясняют, одинаково ли влияние фактора путём исследования значимости между средними в наборах данных, соответствующих градациям фактора.
Минимальное число классов градации (групп) - два. Классы градации могут быть качественными либо количественными.
Применение дисперсионного анализа возможно, если можно предполагать соответствие выборочных групп генеральным совокупностям с нормальным распределением и независимость распределений наблюдений в группах.
Дисперсионный анализ - почти универсальный метод проверки различий в группах.

Однофакторный анализ

Однофакторный дисперсионный анализ основан на сравнении дисперсии между выборочными средними (межгрупповая дисперсия) с дисперсией внутри выборок (внутригрупповая, или случайная, дисперсия). Если межгрупповая дисперсия статистически значимо превосходит внутригрупповую, различия между средними признаются достоверными. При этом считают, что фактор оказывает статистически значимое влияние на исследуемую переменную. Значимость различий проверяется по критерию Фишера.
Примечание: однофакторный дисперсионный анализ можно использовать и для сравнения двух выборочных средних аналогично критерию Стьюдента

В MS Excel однофакторный дисперсионный анализ может быть проведен посредством команды “Однофакторный дисперсионный анализ” из пакета Анализ данных
В результате действия процедуры выводятся две таблицы. Первая таблица - **Итоги**. В ней содержатся данные обо всех классах градации фактора: число наблюдений, суммарное значение, среднее значение и дисперсия.
Во второй таблице - **Дисперсионный анализ** - содержатся данные о величинах для фактора между группами и внутри групп и итоговых. Это сумма квадратов отклонений (**SS**), число степеней свободы (**df**), дисперсия (**MS**). В последних трёх столбцах - фактическое значение отношения Фишера(**F**), p-уровень (**P-value**) и критическое значение отношения Фишера (**F crit**).
Интерпретация результатов. Влияние исследуемого фактора определяется по величине значимости критерия Фишера, которая находится в таблице *Дисперсионный анализ* на пересечении строки *Между группами* и столбца *P-Значение*. В случаях, когда P-Значение < 0,05, критерий Фишера значим и влияние исследуемого фактора можно считать доказанным.

Двухфакторный анализ

Двухфакторный дисперсионный анализ применяется для того, чтобы проверить возможную зависимость результативного признака от двух факторов - *A* и *B*.
Для проведения двухфакторного дисперсионного анализа в пакете Анализ данных реализованы процедуры **Двухфакторный дисперсионный анализ с повторениями** и **Двухфакторный дисперсионный анализ без повторений**.
В результате действия процедуры **Двухфакторный дисперсионный анализ без повторений** выводятся две таблицы. Первая таблица - Итоги. В ней содержатся данные обо всех классах градации фактора: число наблюдений, суммарное значение, среднее значение и дисперсия. Во второй таблице - Дисперсионный анализ - содержатся данные об источниках вариации: рассеивании между строками, рассеивании между столбцами, рассеивании ошибки, общем рассеивании, сумма квадратов отклонений (SS), число степеней

свободы (df), дисперсия (MS). В последних трёх столбцах - фактическое значение отношения Фишера(F), p-уровень (P-value) и критическое значение отношения Фишера (F crit).

Двухфакторный дисперсионный анализ с повторениями применяется для того, чтобы проверить не только возможную зависимость результативного признака от двух факторов - *A* и *B*, но и возможное взаимодействие факторов *A* и *B*. Тогда *a* - число градаций фактора *A* и *b* - число градаций фактора *B*, *r* - число повторений.

В результате действия процедуры выводятся две таблицы. Первая таблица состоит из трёх частей: две первые соответствуют каждой из двух рекламных кампаний, третья содержит данные об обеих рекламных кампаниях. В столбцах таблицы содержится информация обо всех классах градации второго фактора - магазина: число наблюдений, суммарное значение, среднее значение и дисперсия.

Во второй таблице - данные о сумме квадратов отклонений (SS), числе степеней свободы (df), дисперсии (MS), фактическом значение отношения Фишера(F), p-уровне (P-value) и критическом значении отношения Фишера (F crit) для различных источниках вариации: двух факторах, которые даны в строках (выборка) и столбцах, взаимодействии факторов, ошибки (внутри) и суммарных показателях (итого).

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Корреляция — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Корреляционный анализ – это проверка гипотез о связях между переменными с использованием коэффициентов корреляции, он дает возможность точной количественной оценки степени согласованности изменений (варьирования) двух и более признаков.

Коэффициент корреляции (R) – это мера прямой или обратной пропорциональности между двумя переменными.

В случае если изменение одной случайной величины не ведёт к закономерному изменению другой случайной величины, но приводит к изменению другой статистической характеристики данной случайной величины, то подобная связь не считается корреляционной, хотя и является статистической. Значительная корреляция между двумя случайными величинами всегда является свидетельством существования некоторой **статистической связи** в данной выборке, но эта связь не обязательно должна наблюдаться для другой выборки и иметь причинно-следственный характер. Наличие корреляции между двумя результатами, в сущности, означает, что при изменении одного результата другой также изменяется.

Коэффициент корреляции(R). Он может варьировать от –1 до +1:

- R= +1 между признаками существует прямая связь;
- R=–1 между признаками существует обратная связь;
- R=0 связь между признаками отсутствует.

На практике коэффициент корреляции очень редко бывает равен +1 или –1. В биологических исследованиях часто используют следующую (условную) классификацию:

- 0,75<= |R|<= 1 сильная связь;
- 0,50<= |R| < 0,75 умеренная связь;
- |R| < 0,5 слабая связь.

Так как коэффициент корреляции ближе к +1 или -1, он указывает на положительную (+1) или отрицательную (-1) корреляцию между массивами. Положительная корреляция означает, что при увеличении значений в одном массиве значения в другом массиве также увеличиваются. Коэффициент корреляции ближе к 0 означает, что корреляция отсутствует или является слабой.

Следует подчеркнуть, что коэффициент корреляции отражает только *степень линейной связи*. В случае, если есть основания предполагать наличие нелинейной связи, следует воспользоваться средствами регрессионного анализа.

MS EXCEL: КОРРЕЛ(), Корреляция (Анализ данных)

РЕГРЕССИОННЫЙ АНАЛИЗ

Во многих биологических исследованиях возникает необходимость определить, связаны ли между собой изучаемые показатели (рост и вес, содержание азота в почве и содержание нитратов в продукции, уровень органического загрязнения водоема и численность сине-зеленых водорослей и т.д.), а также определить ха-рактер и возможные причины этой связи. Математические методы, основанные на данном подходе, известны под названиями корреляционного и регрессионного анализов. Корреляционный анализ предназначен для изучения линейных зависимостей (зависимостей, которые можно описать уравнением вида y=ax+b).Регрессионный анализ предназначен для изучения любых зависимостей. В определенном смысле корреляционный анализ является частным случаем регрессионного анализа. Изучение взаимозависимостей в MS Excel может быть проведено средствами надстройки **Пакет анализа**, а также **встроенными функциями MS Excel** и посредством **линии тренда**.

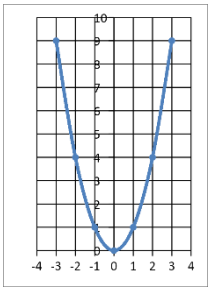
Аппроксимация экспериментальных данных

Известны 3 основных способа задания функциональных зависимостей:

1. Аналитический

y=x2

2. Графический



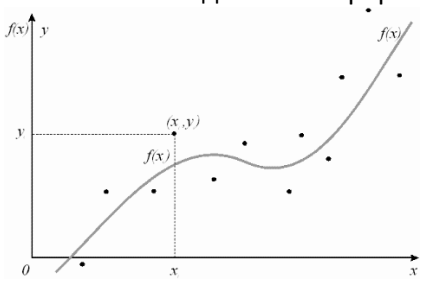
3. Табличный

x	y
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9

Табличный способ обычно возникает в результате эксперимента.

Недостаток табличного задания функции, с которым мы чаще всего встречаемся при решении прикладных задач, заключается в том, что всегда найдутся значения переменных, которые неопределенны таблицей. Для отыскания таких значений определяют **приближающуюся к заданной функцию**, называемой **аппроксимирующей**, а действие замены аппроксимацией (приближением).

Основная задача аппроксимации – **построение приближенной** (аппроксимирующей) функции наиболее близко проходящей около данных точек или около данной непрерывной функции.



> Аппроксимация – процесс подбора эмпирической (полученной опытным путем) функции ф(х) для установления из опыта функциональной зависимости у= ф(х)

Эмпирические формулы служат для аналитического представления опытных данных.

В простейшем случае задача аппроксимации экспериментальных данных выглядит следующим образом:

Пусть есть какие-то данные, полученные практическим путем (в ходе эксперимента или наблюдения), которые можно представить парами чисел (х,у). Зависимость между ними отражает таблица:

На основе этих данных требуется подобрать функцию у=ф(х), которая наилучшим образом сглаживала бы экспериментальную зависимость между переменными и по возможности точно отражала общую тенденцию зависимости между х и у, исключая погрешности измерений и случайные отклонения. Это значит, что отклонения yi-ф(xi) в каком–то смысле были бы наименьшими.

Обычно задача аппроксимации распадается на две части:

1. Сначала устанавливают вид зависимости y=f(x) и, соответственно вид эмпирической формулы, то есть решают, является ли она линейной, квадратичной, логарифмической или какой-либо другой.
2. После этого определяются численные значения неизвестных параметров выбранной эмпирической формулы, для которых приближение к заданной функции оказывается наилучшим.

Обычно определение параметров при известном виде зависимости осуществляют по **методу наименьших квадратов**. При этом функция ф(х) считается наилучшим приближением к f(x), если для нее сумма квадратов невязок $\varepsilon_i=f(x_i)-\phi(x_i)$ или отклонений "теоретических" значений $\phi(x_i)$, найденных по эмпирической формуле, от соответствующих опытных значений y_i, имеет наименьшее значение по сравнению с другими функциями, из числа которых выбирается искомое приближение.

Линия тренда

В MS Excel аппроксимация экспериментальных данных может быть реализована путем построения диаграммы по исходным данным с последующим подбором подходящей аппроксимирующей функции (линии тренда).

Линии тренда - графическое представление направления изменения ряда данных.

Она представляют собой геометрическое отображение средних значений анализируемых показателей, полученное с помощью какой-либо математической функции.

Линия тренда всегда связана с рядом данных, но не представляет данные этого ряда. Она предназначена для отображения тенденций в существующих данных или прогнозов будущих данных.

Типы диаграмм, поддерживающие линии тренда.

- ненормированные плоские диаграммы с областями,
- линейчатые диаграммы,
- гистограммы,
- графики,
- биржевые диаграммы,
- точечные диаграммы,
- пузырьковые диаграммы.

Если линия тренда добавляется к графику, гистограмме, диаграмме с областями или линейчатой диаграмме, она вычисляется согласно **допущению, что значения X равны 1, 2, 3, 4, 5, 6 и т. д**. Такое допущение делается независимо от того, являются ли значения по оси X числовыми или текстовыми. Чтобы вычислить линию тренда на основе числовых значений X, необходимо использовать **точечную диаграмму**.

При добавлении линии тренда на диаграмму Microsoft Office Excel можно выбрать любой из следующих шести различных типов тренда или регрессии: прямые, логарифмические, полиномиальные, степенные и экспоненциальные линии тренда, а также линии тренда с линейной фильтрацией. Тип линии тренда, который следует выбирать, определяется типом имеющихся данных и характером изменения данных во времени.

Типы линий тренда:

1. Линейная — $y=ax+b$

Обычно применяется в простейших случаях, когда экспериментальные данные возрастают или убывают с постоянной скоростью.

2. Полиномиальная — $y=a_0+a_1x+a_2x^2+\ldots+ a_nx^n$, где n до шестого порядка включительно (n<=6), a_i — константы.

Используется для описания экспериментальных данных, попеременно возрастающих и убывающих. Степень полинома определяется количеством экстремумов (максимумов или минимумов) кривой.

3. Логарифмическая — $y=a\ \ln x+b$, где a и b — константы

Функция применяется для описания экспериментальных данных, которые вначале быстро растут или убывают, а затем постепенно стабилизируются.

4. Степенная — $y=bx^a$, где a и b — константы.

Степенная аппроксимация полезна для описания монотонно возрастающей либо монотонно убывающей величины. Данные не должны иметь нулевых или отрицательных значений.

5. Экспоненциальная — $y=be^{ax}$, где a, b — константы.

Применяется для описания экспериментальных данных, которые быстро растут или убывают, а затем постепенно стабилизируются. Часто ее использование вытекает из теоретических соображений.

6. Скользящее среднее.

Использование в качестве приближения скользящего среднего позволяет сгладить колебания данных и таким образом более наглядно показать характер зависимости. Такая линия тренда строится по определенному числу точек (оно задается параметром **Шаг**). Элементы данных усредняются, и полученный результат используется в качестве среднего значения для приближения.

Коэффициент детерминации

Степень близости аппроксимации экспериментальных данных выбранной функцией оценивается **коэффициентом детерминации** (величина достоверности аппроксимации) (**R^2**). Коэффициент детерминации для модели с константой принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется **как соответствие модели данным**. При аппроксимации данных с помощью линии тренда значение величины достоверности аппроксимации рассчитывается приложением MS Excel автоматически. При необходимости полученный результат можно показать на диаграмме.

- Если $R^2 \geq 0{,}95$,то говорят о высокой точности аппроксимации (модель хорошо описывает явление).
- Если $0{,}75 \leq R^2 < 0{,}95$,то говорят об удовлетворительной аппроксимации (модель в целом адекватно описывает явление).
- Если $0{,}5 \leq R^2 < 0{,}75$,то говорят о слабой аппроксимации (модель слабо описывает явление).
- Если $R^2 < 0{,}5$,то говорят, что точность аппроксимации недостаточна и модель требует изменения.

Таким образом, регрессионный анализ может быть проведен посредством линии тренда.

Наиболее распространенный вид регрессионного анализа — **линейная регрессия,** когда находят линейную функцию, которая, согласно определённым математическим критериям, наиболее соответствует данным. Например, в методе наименьших квадратов вычисляется прямая (или гиперплоскость), сумма квадратов между которой и данными минимальна.

ЛИНЕЙН()

Более полные статистические данные для анализа уравнения регрессии можно получить с помощью функции **ЛИНЕЙН**(), которая вводится как формула массива (Ctrl+Shift+Enter) и служит для расчета данных, описывающих **линейной** множественной или парной регрессии на основе метода наименьших квадратов.

ЛИНЕЙН(известные_значения_y; [известные_значения_x]; [конст]; [статистика])

Аргументы функции ЛИНЕЙН

Известные_значения_y. Обязательный аргумент. Множество значений y, которые уже известны для соотношения y = mx + b.

Если массив ***известные_значения_y*** имеет один столбец, то каждый столбец массива ***известные_значения_x*** интерпретируется как отдельная переменная.

Если массив ***известные_значения_y*** имеет одну строку, то каждая строка массива ***известные_значения_x*** интерпретируется как отдельная переменная.

Известные_значения_x. Необязательный аргумент. Множество значений x, которые уже известны для соотношения y = mx + b.

Массив ***известные_значения_x*** может содержать одно или несколько множеств переменных. Если используется только одна переменная, то массивы ***известные_значения_y*** и ***известные_значения_x*** могут иметь любую форму — при условии, что они имеют одинаковую размерность. Если используется более одной переменной, то ***известные_значения_y*** должны быть вектором (т. е. интервалом высотой в одну строку или шириной в один столбец).

Если массив ***известные_значения_x*** опущен, то предполагается, что это массив {1;2;3;...}, имеющий такой же размер, что и массив ***известные_значения_y***.

Конст. Необязательный аргумент. Логическое значение, которое указывает, требуется ли, чтобы константа b была равна 0.

Если аргумент ***конст*** имеет значение ИСТИНА или опущен, то константа b вычисляется обычным образом.

Если аргумент ***конст*** имеет значение ЛОЖЬ, то значение b полагается равным 0 и значения m подбираются таким образом, чтобы выполнялось соотношение y = mx.

Статистика. Необязательный аргумент. Логическое значение, которое указывает, требуется ли вернуть дополнительную регрессионную статистику.

Если ***статистика*** имеет true, **то LINEST** возвращает дополнительную регрессию; в результате возвращается массив **{mn;mn-1,...,m1;b;sen-1,...,se1;seb;r^2;sey; F,df;ssreg,ssresid}**.

Если аргумент ***статистика*** имеет значение ЛОЖЬ или опущен, функция **ЛИНЕЙН** возвращает только коэффициенты m и постоянную

Дополнительная регрессионная статистика.

	A	B	C	D	E	F
1	m_n	m_{n-1}	...	m_2	m_1	b
2	se_n	se_{n-1}	...	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

Величина	Описание
se1,se2,..., sen	Стандартные значения ошибок для коэффициентов m1,m2,...,mn.
seb	Стандартное значение ошибки для постоянной b (seb = #Н/Д, если аргумент ***конст*** имеет значение ЛОЖЬ).

r2	Коэффициент определения. Сравнивает оценочная и фактическая значения y и диапазоны значений от 0 до 1. Если значение 1, то в выборке будет отличная корреляция— разница между предполагаемым значением y и фактическим значением y не существует. Если коэффициент определения — 0, уравнение регрессии не поможет предсказать значение y.
sey	Стандартная ошибка для оценки y.
F	F-статистика или F-наблюдаемое значение. F-статистика используется для определения того, является ли случайной наблюдаемая взаимосвязь между зависимой и независимой переменными.
df	Степени свободы. Степени свободы используются для нахождения F-критических значений в статистической таблице. Для определения уровня надежности модели необходимо сравнить значения в таблице с F-статистикой, возвращаемой функцией **ЛИНЕЙН**.
ssreg	Регрессионная сумма квадратов.
ssresid	Остаточная сумма квадратов. Дополнительные сведения о расчете величин ssreg и ssresid см. в подразделе "Замечания" в конце данного раздела.

Замечания

- Уравнение прямой имеет вид y = mx + b. Если известны значения m и b, то можно вычислить любую точку на прямой, подставляя значения y или x в уравнение. Можно также воспользоваться функцией **ТЕНДЕНЦИЯ**.
- Точность аппроксимации с помощью прямой, вычисленной функцией **ЛИНЕЙН**, зависит от степени разброса данных. Чем ближе данные к прямой, тем более точной является модель **ЛИНЕЙН**. Функция **ЛИНЕЙН** использует для определения наилучшей аппроксимации данных метод наименьших квадратов. Когда имеется только одна независимая переменная x, значения m и b вычисляются по следующим формулам:

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \bar{y} - m\bar{x}$$

где x и y — выборочные средние значения, в частности,
x = **СРЗНАЧ(известные_значения_x)**,
y = СРЗНАЧ(известные_значения_y).

$$R^2 = \frac{\left(n\sum x_i y_i - \sum x_i \sum y_i\right)^2}{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}$$