Основы научных исследований в садоводстве

Л.1. Введение. Статистические параметры выборки. Закономерности случайной вариации. Оценка достоверности статистических параметров

Москва 2018

Понятие о биометрии. Предмет изучения биометрии

- Биометрия – наука о применении математических методов для изучения биологических организмов. Таким образом, предметом изучения биометрии являются метематические методы, используемые для тех или иных суждений о биологических явлениях и процессах

Задачи биометрии

- Задачи биометрии очень разнообразны, постоянно развиваются и меняются в зависимости от применяемых математических методов:

1. Вычисление биометрических характеристик выборки
2. Оценка достоверности выборочных биометрических характеристик, то есть оценка степени их соответствия генеральным биометрическим характеристикам
3. Оценка достоверности различий между выборками по тем или иным признакам
4. Оценка степени влияния тех или иных факторов на признаки выборки
5. Оценка степени сопряженности варьирования признаков
6. Прогнозирование изменения тех или иных признаков в зависимости от изменения других признаков или факторов

История биометрии

- До XVIII века биология развивалась только на основе качественного анализа явлений, то есть, была описательной наукой. В 1899 году Гальтон разработал основы новой науки, названной им биометрией

Предпосылки для внедрения математики в биологию

- Переход от чисто описательного метода к экспериментальному, эксперимент неизбежно требует количественной оценки явлений и процессов
- Возникновение новых биологических наук в XIX веке: физиологии, генетики, радиобиологии и др.
- Развитие агрономии потребовало разработки:

1. Схем опытов для выяснения влияния на урожайность различных факторов
2. Методов математического анализа результатов опытов
3. Способов доказательства достоверности влияния того или иного фактора, впоследствии этого потребовали – медицина, зоология, ботаника и др.
4. Установление факта, что биологическим явлениям свойственны статистические закономерности, обнаруживаемые при изучении совокупностей, но неприложимые к отдельным единицам совокупности. Например, в зоологии и ботанике переход от изучения типичных представителей вида к изучению популяций

- Роль математики и биологии особенно возрасла с развитием теории информации, кибернетики, программирования
- В настоящее время в биологии широко используются не только статистические методы математики, но и дифференциальное и интегральное исчисления, матричная алгебра и другие области
- В различных областях биологии (генетика, теория эволюции, селекция, физиология) ставится задача выражения биологических процессов в математической форме

Понятие о совокупности

- Совокупностью называется всякое множество отдельных, отличающихся друг от друга и в то же время сходных в некоторых существенных отношениях, объектов
- Совокупность составляют различные члены или единицы совокупности. Число единиц совокупности называют *объемом совокупности* (N,n). Единицами совокупности могут быть отдельные организмы (растения) или части организмов (плоды, семена, листья и т.п.)

- Единица совокупности характеризуется определенными признаками. Признак – это то, что характеризует то или иное свойство единицы совокупности. Каждый признак у различных единиц совокупности принимает разные значения, то есть, варьирует
- Различие между единицами совокупности по тому или иному признаку (переменному) называется вариацией или дисперсией (рассеянием)

- Значение признака у той или иной единицы совокупности называют вариантой и обозначают x_i, где *i*- порядковый номер варианты
- Варьирующую величину, то есть, величину, изменяющуюся под влиянием многих случайных причин и принимающую разные значения, называют случайной переменной *x*. Иными словами варианты являются числовыми значениями x
- Однако, несмотря на различия по тем или иным признакам, члены совокупности однородны, то есть, сходны по некоторым важным признакам

- Большая совокупность может состоять из более мелких, частных совокупностей. Например, совокупность растений того или иного вида, состоит из совокупностей популяций, сортов и т.п.
- Наиболее общую совокупность называют генеральной совокупностью. Генеральная совокупность – теоретически бесконечно большая совокупность из всех единиц, которые могут быть к ней отнесены
- На практике приходится иметь дело с выборочными совокупностями

Понятие о переменных

- Анализ данных сильно зависит от того, каков характер вариации изучаемых признаков. Различают два типа вариации:

1. Качественная вариация – признаки имеют очень ограниченный ряд состояний
2. Количественная вариация

- дискретная: различия между отдельными значениями случайной переменной выражаются целыми числами
- непрерывная: различия между отдельными значениями случайной переменной зависят от степени точности измерений (масштаба, интервала) количественного признака

Способы учета признаков – шкалы оценки

- Чтобы оценить значение признака, необходимо выбрать шкалу оценки. Шкала оценки – это способ измерения состояния переменного
- Существуют 3 типа шкал оценки признаков: номинальная, порядковая и интервальная
- Эти шкалы отличаются друг от друга по двум основным свойствам:

1. Наличию или отсутствию правила ранжирования состояний переменного
2. Наличию или отсутствию заданного интервала между состояниями переменного

Номинальная (категориальная) шкала

- Является низшей шкалой оценки состояний переменного
- Номинальные шкалы используют для оценки качественных признаков
- В общем виде к качественным относят такой признак, состояния которого невозможно количественно измерить. Качественные признаки часто называют номинальными или категориальными признаками
- Категоризованными переменными называют переменные, превращенные в номинальные (категориальные)

- Состояние качественного номинального признака называется модальностью. В связи с этим, признаки в выборке могут быть мономодальными (отсутствие вариации), бимодальными (две модальности) и полимодальными
- Исходные данные для анализа номинальных признаков представляют собой наблюдаемые частоты встречаемости модальностей в выборке
- Единственными математическими связями, уместными по отношению к номинальным шкалам, являются тождество и различие состояний признака у изучаемых объектов. Интервал между модальностями не определен
- Характерные особенности:

1. Правило ранжирования модальностей отсутствует
2. Интервал между модальностями не

Порядковая (ранговая, ординальная) шкала

- Применяется для таких переменных, у которых их отдельные состояния можно упорядочить (ранжировать)
- Используют в основном для оценки качественных признаков
- Отдельное состояние порядкового признака обычно называют рангом *(R_i).* В дополнение к тождеству и различию для порядковых шкал используются связи типа больше или меньше
- Характерные особенности:

1. Наличие правила ранжирования состояний переменного
2. Интервал между рангами не определен

- В общем виде рангом R_i наблюдения x_i среди величин $x_1, \ldots x_n$ называют тот порядковый номер, который получит значение x_i при расстановке чисел $x_1, \ldots x_n$ в порядке возрастания или убывания
- В случае равенства x_i для нескольких объектов в выборке, рангом будет среднее арифметическое из соответствующих порядковых номеров итих переменных
- Сумма всех рангов в выборке всегда должна быть равна сумме порядковых номеров

Интервальная шкала

- Является основной шкалой оценки количественных признаков. Отдельное состояние признака в интервальной шкале называется вариантой (x_i)
- Для того, чтобы задать интервальную шкалу, надо определить начальную точку и единицу измерения. Далее при измерении ставят в соответствие каждому объекту число, показывающее, на сколько единиц измерения этот объект отличается от объекта, принятого за начальную точку (например, температура, масса и т.п.)
- Характерными особенностями интервальной шкалы являются:

1. Наличие правила ранжирования состояний переменного
2. Интервал между состояниями переменного определен

Группировка данных при качественной вариации

- Для анализа совокупности необходимо провести группировку вариант у различных единиц совокупности. Наиболее проста группировка при качественной вариации
- При этом подсчитываются число единиц совокупности, обладающих одинаковыми состояниями признака (например, частоты встречаемости тех или иных модальностей) и строится таблица частот встречаемости. При этом частоты выражаются либо абсолютными числами (предпочтительно), либо долями или процентами их встречаемости от объема совокупности
- Сумма всех частот по тому или иному признаку должна быть равна объему совокупности или 100%
- Частным случаем качественной вариации является альтернативная вариация, когда совокупность делится на 2 группы: одна группа характеризуется проявлением признака, другая – его отсутствием

Группировка данных при количественной дискретной вариации

- Вначале определяются минимальное и максимальное значения признака (x_{min};x_{max})
- Затем вычисляется размах изменчивости: разность между максимальным и минимальным значением признака *(lim)*
- Далее подбирается межклассовый интервал (λ) и определяются границы классов
- Далее варианты разносятся по классам, и определяется частота встречаемости того или иного класса (n_i)
- В итоге возникает вариационный ряд – то есть, распределение частот встречаемости всех классов
- Класс, обладающий максимальной частотой, называется модальным
- Вариационный ряд обычно изображается графически в виде кривой распределения или вариационной кривой.

- Существует 2 способа графического изображения вариационных рядов:

1. По оси абсцисс наносятся середины классов (среднее значение из всех вариант того или иного класса); по оси ординат – частоты их встречаемости; высота класса, пропорциональная его частоте, отмечается точкой; ломаная линия, соединяющая все точки, называется полигоном распределения; слева и справа полигон распределения должен пересекать ось абсцисс (нулевые классы)
2. по оси абсцисс наносят границы классов (например, минимальные значения вариант в классе): по оси ординат – частоты их встречаемости, высота класса, пропорциональная его частоте, отмечается отрезком, то есть каждый класс выглядит как прямоугольник; совокупность всех прямоугольников называется гистограммой распределения

Рис. 1. Полигон распределения численности поросят в 64 опоросах свиноматок

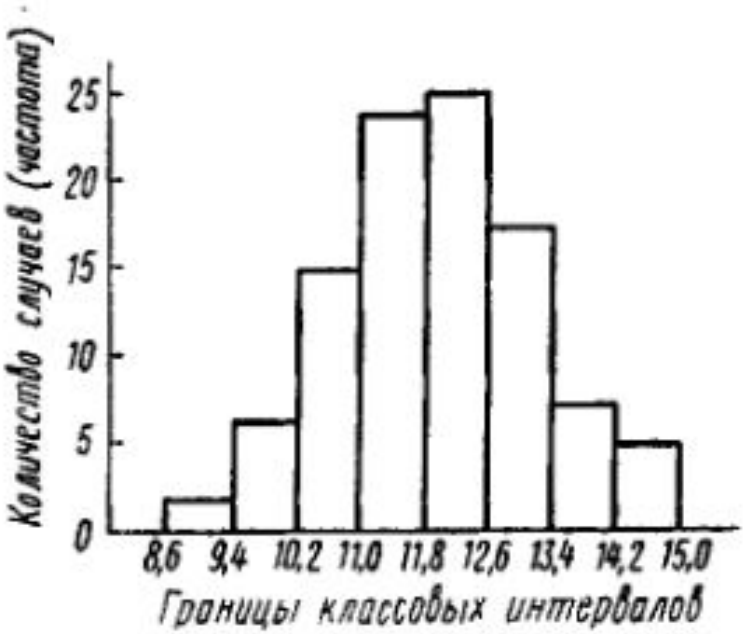


Рис. 2. Гистограмма распределения кальция (мг%) в сыворотке крови обезьян

Группировка данных при количественной непрерывной вариации

- Группировка данных при непрерывной изменчивости наиболее сложная. Основная сложность заключается в разбивке совокупности на классы, так как естественные классы отсутствуют (при качественной изменчивости эти классы представлены явно в виде модальностей, при дискретной изменчивости их определить относительно просто).
- Алгоритм группировки такой же, как и при дискретной изменчивости:
- 1) определение минимального и максимального значений признака (x_{min}, x_{max});
- 2) вычисление размаха изменчивости (lim);
- 3) подбор межклассового интервала (λ);
- 4) разбивка совокупности на классы: определение левой и правой границ каждого класса (левая граница первого класса равна x_{min}, правая граница последнего класса равна x_{max};
- 5) разноска вариант по классам;
- 6) определение частот встречаемости классов (n_i);
- 7) построение полигона частот или гистограммы распределения

Закономерности распределения вариант в вариационном ряду

- Общие закономерности:
- 1) большинство вариант располагается в средней части ряда;
- 2) распределение вариант более или менее симметрично относительно середины;
- 3) к краям вариационного ряда частота убывает

Две группы статистических показателей совокупности

- Вариационные ряды различаются по двум свойствам: 1) по средней тенденции, вокруг которой варьируют варианты; 2) по степени вариации вариант, то есть, по степени отклонения вариант от средней тенденции.
- Соответственно статистические показатели делятся на 2 группы: 1) показатели *средней тенденции*: мода, медиана, средняя арифметическая; 2) показатели *вариации*: размах вариации, среднее абсолютное отклонение, среднее квадратическое отклонение, дисперсия (варианса)

Мода

- *Мода* (Мо) – значение модального класса, то есть класса, который встречается с максимальной частотой. Для количественных признаков Мо – среднее значение (середина) модального класса. Число мод в выборке не определено. Максимальное число мод в выборке может быть равно числу классов, когда все классы встречаются с максимальной частотой. Моду можно вычислить для любого признака, как качественного, так и количественного

Медиана

- *Медиана* (Me) – это значение варианты, которая находится точно в середине (центре) ранжированного вариационного ряда. Для того чтобы определить медиану вначале необходимо ранжировать (упорядочить) варианты от минимальных их значений до максимальных.
- Если объем выборки является четным числом, то медиана является средним значением двух соседних срединных вариант. Если объем выборки является нечетным числом, то медиана является значением срединной (центральной) варианты.
- Свойства медианы: 1) медиана в выборке всегда одна; 2) медиана относительно устойчива, и наименее зависит от значений отдельных вариант.
- Медиану можно вычислить только для признаков, оцененных в порядковой или в интервальной шкалах. Если признак оценен в номинальной шкале, медиану определить невозможно

Среднее арифметические

Среднее арифметическое ($\bar{x}$) представляет собой частное от деления суммы всех вариант выборки ($\sum_{i=1}^{N} x_i$) на объем выборки (N):

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Обычно среднее арифметическое вычисляется на один знак после запятой точнее, чем отдельные наблюдения.
- Свойства средней арифметической:
- 1) если каждую из вариант совокупности увеличить или уменьшить на одну и ту же величину, то и средняя арифметическая соответственно уменьшится или увеличится на эту же величину;
- 2) сумма разностей между отдельными вариантами и средней арифметической равна нулю: $\Sigma(x_i - \bar{x})=0$;
- 3) сумма квадратов отклонений вариант от средней арифметической $\Sigma(x_i - \bar{x})^2$ всегда меньше суммы квадратов отклонений вариант от любой другой величины «А» не равной средней арифметической: $\Sigma(x_i - \bar{x})^2 < \Sigma(x_i - A)^2$, если $A \neq \bar{x}$.
- Особенности средней арифметической:
- 1) средняя арифметическая характеризует всю совокупность в целом, а не отдельные единицы совокупности;
- 2) средняя арифметическая имеет смысл только по отношению к качественно однородной совокупности;
- 3) средняя арифметическая характеризует только данную совокупность, экстраполировать её рискованно.
- 4) средняя арифметическая вычисляется только для признаков, измеренных в интервальной шкале

Размах изменчивости

- Размах изменчивости (*lim*) – разница между максимальным и минимальным значениями признака в совокупности:

$$\lim = x_{\max} - x_{\min}$$

- Недостатки данного показателя: 1) очень не устойчивый (зависит только от крайних значений совокупности); 2) при равенстве размаха изменчивости двух выборок, распределение в них вариант может быть разным

Дисперсия

- *Дисперсия* (варианса, σ^2) в общем виде – это средний квадрат отклонений вариант от средней арифметической сово

Для *генеральной совокупности* – варианса ($\hat{\sigma}^2$) – это частное от деления суммы квадратов отклонений отдельных вариант от средней арифметической $\Sigma(x_i - \bar{x})^2$ на объем выборки N:

$$\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

Для *выборочной совокупности* – варианса (σ^2) - это частное от деления суммы квадратов отклонений отдельных вариант от средней арифметической (SS):

$$SS=\Sigma(x_i - \bar{x})^2$$

на число степеней свободы (df):

$$df=N-1$$

$$\sigma^2 = \frac{SS}{df} = \frac{\sum(x_i - \bar{x})^2}{N-1}$$

Число степеней свободы в общем виде – это число независимых наблюдений при оценке дисперсии.

Выше приведенные формулы вычисления дисперсии смысловые, расчеты проводить по ним очень неудобно.

Рабочая формула для вычисления дисперсии следующая:

$$\sigma^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N-1}$$

Дисперсия является неименованной величиной, то есть, не измеряется в каких-либо единицах, поскольку представляет собой квадрат.

Среднее квадратическое отклонение

- Представляет собой корень квадратный из дисперсии:

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

- В отличие от дисперсии измеряется в тех же единицах, что и признак. Среднее квадратическое отклонение может быть как положительным, так и отрицательным числом (±σ)

Коэффициент вариации

- Коэффициент вариации (*cv*) применяется для сравнения вариации разных признаков. Коэффициент вариации есть частное от деления среднего квадратического отклонения (σ) на среднюю арифметическую:

$$cv = \frac{\sigma}{\bar{x}} \times 100\%$$

- Обычно выражается в процентах. Отличается относительной устойчивостью. Прямо пропорционален среднему квадратическому отклонению и обратно пропорционален среднему арифметическому

Основные статистические параметры выборки

- 1) **объем выборки** (*N*);
- 2) **среднее арифметическое** ($\bar{x}$) как наиболее важный показатель средней тенденции;
- 3) **дисперсия** (σ^2) как основной показатель вариации

Закономерности случайной

Понятие о вероятности и статистической закономерности

- Отдельные члены совокупности, как правило, варьируют. Каждый из них представляет собой как бы отдельный случай, который осуществляется под влиянием многих определяющих причин.
- То есть, каждое отдельное явление, взятое само по себе, представляется случайным (например, длина отдельного листа на дереве), но, взятые в массе, они обнаруживают определенные, так называемые *статистические закономерности*. В отношении же каждого единичного явления приходиться говорить лишь только об известной возможности, или вероятности, значения, которое они приобретают

- *Вероятность* – это возможность осуществления определенного события в некотором количестве случаев из общего числа возможных. Другими словами, вероятность – это степень уверенности в том, что событие произойдет. Процесс осуществления явления (события) на основе его вероятности называется вероятностным или стохастическим процессом.
- Математически – вероятность (p), есть частное от деления числа благоприятных случаев (m) на число всех равновозможных случаев (N). Вероятность варьирует от 0 до 1. При приближении к нулю событие произойдет с малой вероятностью, то есть, в среднем, очень редко. При приближении к единице, наоборот, с большой вероятностью, то есть, почти всегда

Эмпирическая и теоретическая вероятности

- Эмпирические вероятности приложимы только к конкретным совокупностям, для которых они вычислены. По эмпирическим вероятностям можно судить о теоретических (априорных) вероятностях
- В генеральной (стохастической) совокупности вероятности становятся теоретическими. Возникает вопрос о том, насколько достоверны статистические показатели, полученные по выборочной совокупности, чтобы можно было по ним судить о генеральной совокупности

Распределение вероятностей

- Вариационный ряд с характерным для него расположением большинства вариант вблизи его центральной части и рассеиванием к краям ряда является в то же время и распределением вероятностей.
- Следовательно, случайная переменная «х» принимает разные значения: $x_1, x_2, x_3, \ldots, x_n$
- под влиянием разнообразных и независимых причин, то есть, её вариация случайная.
- Отдельным значениям x_i можно придать соответствующие вероятности: $p_1, p_2, p_3, \ldots, p_n$
- Совокупность значений x_i и соответствующих им вероятностей p_i и называется *распределением*

Если вероятности появления отдельных значений x_i выражаются величинами, соответствующими коэффициентам разложения бинома Ньютона, распределение называется биномиальным:

$(a+b)^1=a+b$	1:1
$(a+b)^2=a^2+2ab+b^2$	1:2:1
$(a+b)^3=a^3+3a^2b+3ab^2+b^3$	1:3:3:1
$(a+b)^4=a^4+4a^3b+6a^2b^2+4ab^3+b^4$	1:4:6:4:1

Биномиальное распределение относится к признакам, варьирующим дискретно, прерывисто. Частоты отдельных классов распределения пропорциональны коэффициентам разложения бинома Ньютона:

$$(p+q)^k$$

где p - вероятность появления данного события, q – вероятность непоявления, k - число классов.

При биномиальном распределении средняя арифметическая и среднее квадратическое отклонение вычисляются по следующим формулам:

$$\bar{x} = kp$$

$$\sigma = \sqrt{kpq}$$

Нормальное распределение

- При биномиальном распределении значение показателя степени «k» бинома $(p+q)^k$ конечно. При приближении «k» к бесконечности распределение становится непрерывным. Полигон распределения превращается в симметричную кривую, которая называется нормальной вариационной кривой. Само же распределение называется нормальным.
- Очень многие биологические характеристики с непрерывной вариацией приближаются к нормальному распределению.
- Теоретическая основа вариации та же, что и при биномиальном распределении: вариация в совокупности – результат совместного действия многих разнонаправленных и независимых друг от друга факторов.
- Согласно теореме М.М.Ляпунова, если случайная величина является суммой большого числа независимых слагаемых, то она с достаточной степени точности будет распределяться по нормальному закону. Поэтому закон нормального распределения – один из основных законов статистики.

- Для изучения вариации при нормальном распределении широко пользуются так называемым *нормированным отклонением* (t): отклонение варианты от среднего арифметического, выраженное в сигмах

$$t = \frac{(x_i - \bar{x})}{\sigma}$$

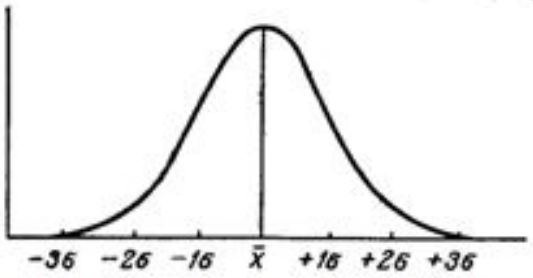


Рис.3. Кривая нормального распределения

- Отсюда следует, что:

$$(x_i - \bar{x}) = t\sigma$$

- Закономерности нормального распределения:
- 1) распределение абсолютно симметрично;
- 2) отклонения вариант от средней арифметической охватывают приблизительно 6σ: 3σ вправо от средней и 3σ влево от средней (рис.3);
- 3) в пределах ±1σ располагается 68,3% всех вариант ряда, в пределах ±2σ - 95,5%, в пределах ±3σ – 99,7% всех вариант;
- 4) значения t для отдельных вариант колеблются в пределах примерно ±3;
- 5) вероятность любого отклонения от средней есть функция нормированного отклонения

- Имеется специально составленная таблица так называемого нормального интеграла вероятностей. Геометрически величины в этой таблице являются долями площади нормальной кривой в границах от –t до +t. Эти доли выражают в то же время и вероятность.
- Закономерности нормального распределения дают возможность по среднему арифметическому и среднему квадратическому отклонению построить весь ряд.
- Если известен размах изменчивости, то его шестая часть приблизительно будет равна σ, а среднее арифметическое будет равно сумме минимального значения варианты и 3σ

Доверительные вероятности

В биометрии существенно важны 2 вероятности: 0,95 и 0,99. Эти вероятности получили название доверительных вероятностей.

С вероятностью 0,95 любая случайно взятая варианта будет отклоняться от среднего арифметического не более чем на ±1,96σ, иными словами, с вероятностью 0,05 варианта будет за пределами ±1,96σ.

С вероятностью 0,99 любая случайно взятая варианта будет отклоняться от среднего арифметического не более чем на ±2,58σ, иными словами, с вероятностью 0,01 варианта будет за пределами ±2,58σ.

Доверительная вероятность 0,99 соответствует более высокому уровню вероятности, соответственно 0,95 более низкому.

Доверительные вероятности определяют доверительные границы и доверительный интервал между ними:

Доверительная вероятность	Доверительный интервал
0,95	-1,96 σ…+1,96σ
0,99	-2,58σ…+2,58σ

Уровни значимости

- Определенным значениям вероятностей соответствуют так называемые уровни значимости.
- Вероятности 0,95 (95%) соответствует уровень значимости 0,05 (5%). Это означает, что выход за пределы принятых границ возможен с вероятностью 0,05, то есть, вероятность ошибочного прогноза составляет 5%.
- Вероятности 0,99 (99%) соответствует уровень значимости 0,01 (1%). Это означает, что выход за пределы принятых границ возможен с вероятностью 0,01, то есть, вероятность ошибочного прогноза составляет 1%.
- Уровень значимости 0,01 более высокий, 0,05 – более низкий.
- Наивысший уровень значимости 0,001 (0,1%) соответствует доверительной вероятности 0,999 (99,9%)

Проблема достоверности в статистике

- Проблема достоверности состоит в расхождении между статистическими показателями выборки и статистическими показателями генеральной совокупности. Если статистические показатели выборки близки к статистическим показателям генеральной совокупности – их достоверность считается высокой. Если статистические показатели выборки сильно отличаются от показателей генеральной совокупности – они недостоверны.
- Выборка должна быть *репрезентативной*. То есть она должна формироваться на основе случайного отбора вариант. Существуют специальные методы позволяющие оценить степень репрезентативности выборки

Ошибка репрезентативности средней арифметической

Средняя арифметическая выборки обозначается символом $\bar{x}$. Средняя арифметическая генеральной совокупности обозначается μ. Каково соотношение между $\bar{x}$ и μ?

Распределение выборочных средних вокруг генерального среднего близко к нормальному закону. Среднее квадратическое отклонение выборочных средних вокруг генерального называется *средней ошибкой* (средней квадратической ошибкой, стандартной ошибкой, ошибкой выборочности, ошибкой репрезентативности). Чем меньше эта ошибка, тем ближе выборочное среднее к генеральному среднему.

Ошибка средней (m_x) вычисляется по формуле:

$$m_x = \frac{\sigma}{\sqrt{N}}$$

Следовательно, величина средней ошибки прямо пропорциональна среднему квадратическому отклонению и обратно пропорциональна объему выборки. Средняя ошибка - это статистическая ошибка, и не имеет ничего общего с ошибкой точности измерений признака.

Чем больше объем выборки, тем меньше разница между средней арифметической выборки и средней арифметической генеральной совокупности. При $N=\infty$, ошибка средней становится равной нулю, то есть, средние арифметические выборки и генеральной совокупности оказываются равными. В этом состоит теорема П.Л.Чебышева.

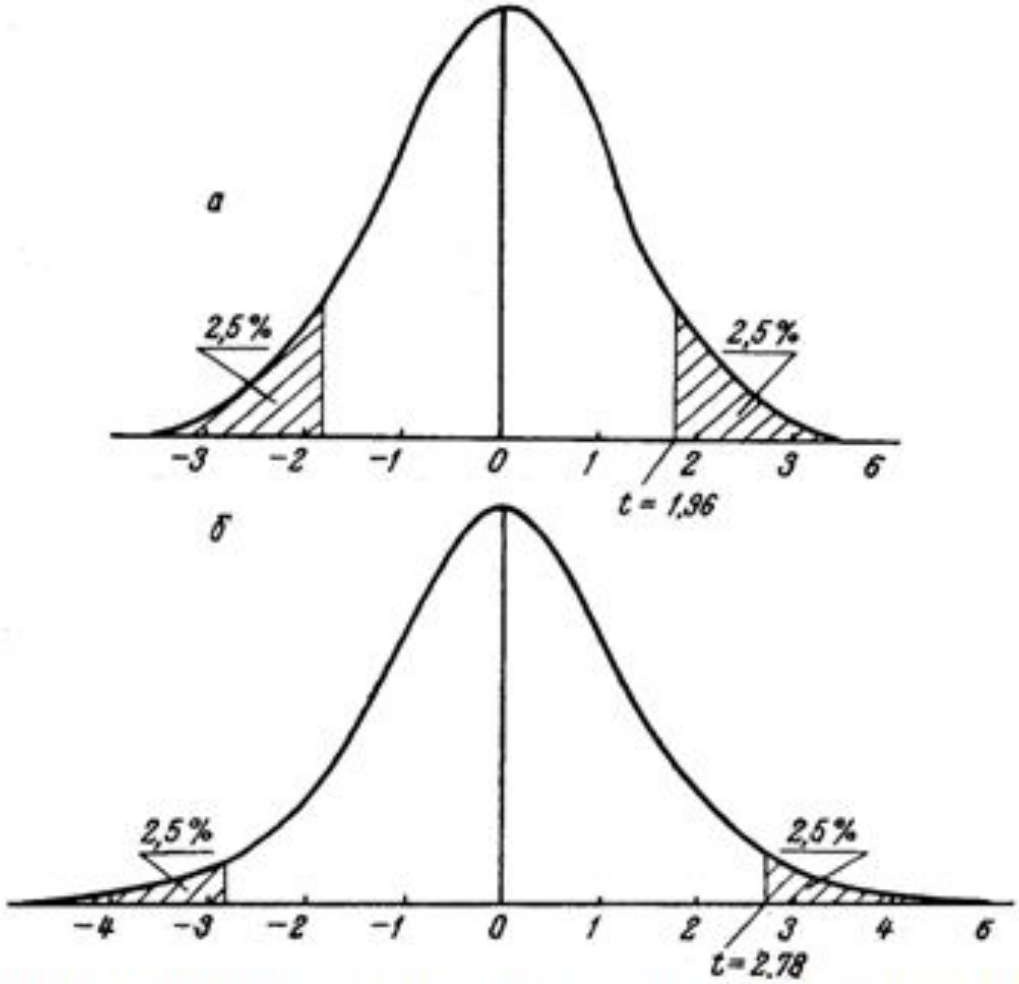
Распределение средних арифметических малых выборок

В больших выборках распределение средних арифметических близко к нормальному закону. Если выборки малы ($N<30$), возникает вопрос о достоверности их статистических параметров.

Ответ на этот вопрос дал английский математик Госсет (псевдоним Стьюдент). Описанное им распределение вероятностей получило название t-распределения по Стьюденту. Критерий t по Стьюденту представляет собой следующее:

$$t = \frac{(\bar{x} - \mu)}{m_x}$$

Распределение значений t отличается от нормального (крутизной кривой) тем сильнее, чем меньше значение N. По мере увеличения «N» t-распределение приближается к нормальному. При $N>30$ разница между распределением Стьюдента и нормальным распределением отсутствует



Значения коэффициента Стьюдента: а – под кривой нормального распределения (N=∞, t=1,96), б – под кривой t-распределения по Стьюденту (N=5, t=2,78)

Доверительный интервал средней арифметической генеральной совокупности

Для определения *доверительного интервала* μ необходимо знать: среднюю арифметическую выборки ($\bar{x}$), ошибку средней (m_x) и критерий Стьюдента (t_{st}) определенного уровня значимости (t_{01} или t_{05}):

$$\bar{x} - tm_x \le \mu \le \bar{x} + tm_x$$

Таким образом, доверительный интервал генеральной средней это её амплитуда варьирования: от минимума до максимума.

Если $N>30$ вместо значения t берётся нормальный интеграл вероятности, если $N<30$ берется t из таблицы Стьюдента.

Определение необходимого объема выборочной совокупности

- Для определения необходимого объема выборки необходимо задать следующие параметры: 1) желаемую точность (Δ) – допустимое расхождение между средней арифметической выборки и средней арифметической генеральной совокупности; 2) коэффициент Стьюдента для определенной доверительной вероятности (если p=0,95, то t_{st} обычно берется равное 2); 3) среднее квадратическое отклонение (σ):

$$N = \frac{t_{st}^2 \sigma^2}{\Delta^2}$$

Нулевая гипотеза

- Общие принципы сравнения выборок основываются на анализе так называемой нулевой гипотезы (H_0). Согласно этой гипотезе, первоначально принимается, что между показателями разных выборок достоверного различия нет. Задача статистического анализа заключается либо в принятии нулевой гипотезы, либо в её отклонении.
- Отбрасывание или принятие нулевой гипотезы связано с принятием того или иного уровня достоверности утверждений (значимости).
- Существует и противоположная нулевой – альтернативная гипотеза (H_1), смысл которой противоположен

Оценка достоверности различий между выборочными средними арифметическими

- Разница между средними арифметическими генеральных совокупностей всегда достоверна, даже если она очень мала, поскольку эти средние были вычислены для генеральных совокупностей.
- Другое дело, если сравниваются две выборочные совокупности. В этом случае необходимо доказывать, что разница между средними арифметическими достоверна.
- Для установления достоверности разницы между средними арифметическими используют нормированные отклонения (t):

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{m_{x_1}^2 + m_{x_2}^2}} = \frac{d}{m_d}$$

- Нулевая гипотеза отвергается, если разница между средними арифметическими достоверна, то есть, критерий Стьюдента будет больше стандартного значения (t_{st}) при определенном уровне значимости.
- В противном случае выборочные средние достоверно не отличаются друг от друга и нулевая гипотеза принимается

Сравнение средних квадратических отклонений и дисперсий

- Сравнение вариансов проводится с использованием критерия Фишера, представляющего отношения дисперсий. Алгоритм вычисления связан с проведением дисперсионного анализа

Основы дисперсионного анализа. Однофакторный дисперсионный анализ

Задачи дисперсионного анализа

- Сущность дисперсионного анализа заключается в установлении влияния отдельных факторов на изменчивость того или иного признака. Сложность анализа состоит в том, что на признаки влияют многочисленные случайные факторы, не поддающиеся контролю
- В связи с этим возникает задача разложения общей вариации (дисперсии) признака на составные элементы, часть из которых определяется изучаемыми конкретными факторами (или фактором), а часть – случайными причинами. Дисперсионный анализ позволяет оценить значимость и долю влияния отдельных факторов и их взаимодействия на вариацию того или иного признака. И, наконец, дисперсионный анализ позволяет оценить достоверность различий между средними по градациям факторов
- Дисперсионный анализ был разработан английским математиком и биологом Р.Фишером

Общие теоретические предпосылки анализа

- Предположим, что на изменчивость какого либо признака оказывает влияние какой-то один фактор. Например, на урожайность растений – дозы внесения удобрений.
- Если рассматривать отклонение отдельного переменного (урожайности) от среднего, то в этом отклонении фигурируют два компонента: 1) отклонение, зависящее от данного фактора (удобрения); 2) остаточная часть, не зависящая от данного фактора:

$$x_i - \overline{x} = A + e$$

- Где, А – доля отклонений переменной, связанная с влиянием данного фактора; е – остаточная часть отклонения (результат случайных отклонений).
- Приведенную схему можно перенести на общую вариацию всех наблюдений, то есть, выразить её в дисперсиях:

$$\sigma_y^2 = \sigma_A^2 + \sigma_e^2$$

- Где σ_y^2 - общая варианса признака; σ_A^2 - варианса, определяемая фактором А; σ_e^2 - варианса, определяемая случайными причинами

- Усложним задачу: на признак оказывают влияние 2 фактора А и В. Например, на ту же урожайность – дозы внесения удобрений и площадь питания растений. Тогда:

$$x_i - \bar{x} = A + B + AB + e$$

- Где А – доля отклонения, связанная с влиянием фактора А; В – доля отклонения, связанная с влиянием фактора В; АВ – доля отклонения, связанная со взаимодействием двух факторов; е – случайная часть отклонения.
- В значениях варианс общая дисперсия будет следующей:

$$\sigma_y^2 = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_e^2$$

Градации факторов

- Для того, чтобы влияние фактора можно было изучить, этот фактор должен иметь несколько состояний или уровней. Эти состояния и называют градациями фактора
- В пределах той или иной градации фактора отдельные переменные варьируют под влиянием случайных причин (случайная вариация)
- Градации факторов могут быть разных типов: 1) фиксированные, например, год наблюдения, месяц, район возделывания, сорт и т.д.; 2) случайные, например, число растений в семье и т.п.

Схемы дисперсионного анализа

- Схемы дисперсионного анализа различаются по следующим особенностям:
- 1) по числу факторов – однофакторные, двухфакторные и т.д.;
- 2) по типу градаций факторов – с фиксированными градациями, со случайными градациями, со смешанными (одни факторы – с фиксированными, другие – со случайными;
- 3) по сочетанию градаций разных факторов – полные (градации одного фактора сочетаются с каждой градацией другого фактора) и иерархические (градации одного фактора связаны с градациями другого фактора по иерархической схеме);
- 4) по числу наблюдений по каждой градации фактора – равномерные (число наблюдений одинаковое) и неравномерные (число наблюдений неодинаковое).

Ограничения

- При проведении дисперсионного анализа должны соблюдаться следующие правила:
- 1) число градаций по фактору должно быть не менее двух;
- 2) число наблюдений по сочетанию градаций разных факторов должно быть не менее двух;
- 3) дисперсии по градациям факторов должны быть примерно одинаковыми;
- 4) распределение величин по градациям факторов должно соответствовать нормальному распределению

Нулевая гипотеза

- Нулевая гипотеза во всех схемах дисперсионного анализа состоит в том, что вся вариация признака является только случайной и не зависит от влияния тех или иных факторов
- Альтернативная гипотеза состоит в признании влияния того или иного фактора или взаимодействия факторов на изменчивость признака

Общие этапы дисперсионного анализа

Алгоритм вычислений включает следующие шаги:

1) вычисление сумм квадратов отклонений (SS);
2) вычисление чисел степеней свободы (df);
3) вычисление средних квадратов (ms);
4) вычисление критерия Фишера (F);
5) определение критических значений критерия Фишера (F_{05} и F_{01});
6) определение достоверности влияния факторов;
7) вычисление дисперсий (σ^2);
8) вычисление долей влияния факторов (p^{ω});
9) построение таблицы результатов дисперсионного анализа;
10) вычисление наименьшей существенной разности (НСР) между средними;
11) сравнение групповых средних.

Спасибо за внимание!!!