МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ЗЕМЛЕУСТРОЙСТВЕ

Карпиченко Александр Александрович

доцент кафедры почвоведения и геоинформационных систем

Всего 46 ч., в т.ч. лекций 24 ч., лабораторных 18 ч., УСР 4 ч.

1.3. Показатели описательной статистики

Показатели рассеивания вариант

Для характеристики распределения в вариационном ряду недостаточно лишь средней арифметической. В совершенно разных по величине вариантах двух выборок средняя может быть одной и той же:

–100; –20; 100; 20; M = 0,

0,1; –0,2; 0,1; M = 0.

Для получения более полного представления о выборочных совокупностях используют показатели рассеяния вариант, или разнообразия признаков: *лимит, размах варьирования (амплитуда),среднеквадратическое (стандартное) отклонение, средний квадрат отклонений (дисперсия), коэффициент вариации, квантили*. Эти показатели признаков характеризуют различную степень и особенности разброса.

1.3. Показатели описательной статистики

Лимит указывает границы вариационного ряда: $Lim = x_{max} \div x_{min}$.

Амплитуда (вариационный размах, размах варьирования) – разность между максимальным и минимальным значениями вариант: $Ampl = x_{max} - x_{min}$.

Чем ближе минимальные и максимальные варианты к среднему и чем меньше амплитуда, тем меньше степень разнообразия между переменными в вариационном ряду, тем надежнее характеризуют статистические показатели искомую закономерность.

1.3. Показатели описательной статистики

Более точно степень разнообразия признака следует характеризовать другими показателями. Среднеквадратическое отклонение и дисперсию используют как составляющие параметры нормального распределения при вычислении ряда параметрических статистических критериев.

Среднеквадратическое отклонение, или сигма (σ) показывает степень рассеяния значений статистической совокупности около среднего значения, а точнее, интервал (M±σ), в который входит до 75 % вариант выборочной совокупности. Считается, если 75 % вариант выборки находится в пределах M±σ. то это соответствует норме (стандартному отклонению); если в пределах M±2σ , то имеется незначительное отклонение от нормы; если выходит за пределы M±3σ, то можно утверждать о наличии анормального явления, процесса. Величина сигмы прямо пропорционально зависит от разброса вариант в вариационном ряду. Чем больший разброс, тем больше значение сигмы. Однако пределы колебаний не дают оценки разброса, как и дисперсия независимо от его величины

1.3. Показатели описательной статистики

Среднеквадратическое отклонение (σ) используется для:

- оценки данных одноименных вариационных рядов при близких средних: чем больше сигма, тем больший разброс вариант в совокупности, соответственно среднее арифметическое менее типично для данного ряда;
- для оценки типичности среднего арифметического в ряду, используя правило трех сигм (M±3σ);
- для определения доверительных интервалов статистических коэффициентов и репрезентативности выборочных исследований.

Недостаток сигмы, как и дисперсии, заключается в том, что критерий представляет собой абсолютную именованную величину, поэтому его нельзя использовать при сравнении разнородных рядов, выраженных в различных единицах измерения. Для этой цели подходит коэффициент вариации.

1.3. Показатели описательной статистики

Среднеквадратическое отклонение можно определить двумя путями:

$$\sigma = \sqrt{\sum (x_i - M_x)^2 / (N-1)} \qquad (1.5);$$

$$\sigma = (x_{max} - x_{min}) / 6 \qquad (1.6),$$

где $(x_i - M_x)$ – отклонение от среднего индивидуальных вариант; N – объем выборочной совокупности.

1.3. Показатели описательной статистики

Средний квадрат отклонений, или дисперсия, указывает колебание значений признака внутри выборочной совокупности через отклонение всех вариант от среднего значения, т. е. показывает интервал, в который входят все варианты выборки (100 %). Однако сумма всех отрицательных и положительных отклонений от среднего равна нулю. Поэтому все отклонения от среднего возводятся в квадрат и суммируются: $\sum(x_i - M_x)^2$. При усреднении всех отклонений числового ряда путем деления на (N – 1) получаем средний квадрат отклонений, или дисперсию (D, σ^2).

Если вычислена сигма (σ), то дисперсию получаем путем возведения ее в квадрат: σ^2.

1.3. Показатели описательной статистики

Если вычислена сигма (σ), то дисперсию получаем путем возведения ее в квадрат: σ^2.

При упрощенном способе расчета дисперсии не вычисляют отклонений вариант от среднего ($x_i - M_x$), используя следующий расчет:

$$\sigma^2 = \sum x_i^2 / N - M^2,$$

где $\sum x_i^2$ – сумма квадратов всех вариант выборки; M^2 – квадрат среднего арифметического; N – число вариант в выборке.

Более точно значение дисперсии вычисляется с использованием данных в табл. 1.3 по формуле:

$$\sigma^2 = \sum (x_i - M_x)^2 / (N - 1)$$

1.3. Показатели описательной статистики

При объединении нескольких аналогичных выборок в общую выборку можно рассчитать общий средний квадрат отклонений, если имеются сведения о дисперсии по каждой выборке в отдельности:

$$\sigma^2_{общ} = \sum(N_i - 1) \cdot \sigma^2_i / (\sum N_i - k), \qquad (1.8)$$

где σ^2_i – дисперсия индивидуальной выборки; N_i – объем частных выборок; k – число частных выборок.

Практическое применение дисперсии (σ^2) состоит в следующем:

- для оценки вариабельности рядов распределения;
- для факторного и дисперсионного анализа;
- для статистической оценки двух совокупностей по критерию Фишера.

Дисперсия выражается в тех же единицах, что и показатели выборки.

1.3. Показатели описательной статистики

Коэффициент вариации представляет собой относительный показатель разнообразия признаков, выражается в процентах. Он показывает отношение среднеквадратического отклонения к средней арифметической:

$$V = (\sigma / M) \cdot 100. \qquad (1.9)$$

В случаях, когда значение среднеквадратического отклонения не рассчитывается, величина коэффициента вариации может быть определена следующим образом:

$$V = 100 \sqrt{\frac{\sum x_i^2 / (M^2 - N)}{N - 1}}, \qquad (1.10)$$

где $\sum x_i^2$ – сумма квадратов индивидуальных вариант в совокупности.

1.3. Показатели описательной статистики

Чем меньший по размаху варьирования будет признак, тем меньший будет коэффициент вариации для данной совокупности. Соответственно меньшими будут сигма и дисперсия.

Коэффициент вариации позволяет оценить вариабельность (разброс) признака в нормированных границах. Если его значение меньше 10 %, то разброс вариант относительно средней арифметической считается слабым, при 10–30 – средним, 30–60 – высоким, 60–100 – очень высоким, более 100 % – аномальным.

При наличии вариант признака с отрицательным числом (отрицательные температуры, отметка поверхности ниже уровня воды в океане и др.) коэффициент вариации рекомендуется вычислять по формуле с учетом модуля:

$$V = 100\ \sigma / |x_i| + M, \qquad (1.11)$$

где $|x_i|$ – модуль наименьшей отрицательной величины без учета знака.

1.3. Показатели описательной статистики

О преимуществе использования коэффициента вариации при оценке разнородных признаков можно судить по таблице, где абсолютные величины средних и сигмы близки по стажу работы и образованию. Однако по коэффициенту вариации сходны по возрасту и образованию. В данном случае сравнение по сигме проводить не корректно, так как все три признака разнородны и не сравнимы между собой. Выручает неименованный коэффициент вариации, который позволяет оценить разброс признака в нормированных границах.

Сравнительная оценка состава работников предприятия

Учетный признак	Среднее арифметическое, *M*	Среднеквадратическое отклонение, σ	Коэффициент вариации, *V*
Стаж работы (лет)	8,7	2,8	32,1
Возраст (лет)	37,2	4,1	11,0
Образование (класс)	9,2	1,1	11,9

1.3. Показатели описательной статистики

Квантили. В открытых вариационных рядах и рядах распределения качественных признаков для сжатого описания распределений используется другой параметр разброса – *квантиль* (синонимы: перцетиль, персентиль). Этот параметр может использоваться для перевода количественных признаков в качественные. В практике статистического анализа наиболее часто используются следующие квантили:

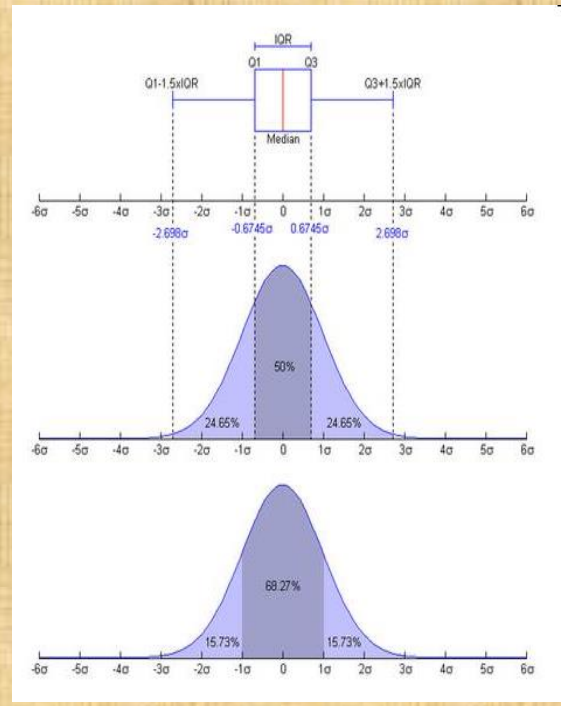
$V_{0,5}$ – медиана;

$V_{0,25}$, $V_{0,75}$ – квантили четверти, соответственно нижняя и верхняя квантиль;

$V_{0,1}$, $V_{0,2}$, … $V_{0,9}$ – децили (десятые);

$V_{0,01}$, $V_{0,02}$, … $V_{0,99}$ – процентили, или центили (сотые).

Квантили делят область возможных изменений вариант в выборке на определенные интервалы. Статистическая суть квантилей лучше раскрывается при построении графика.



1.4. Оценка статистических параметров по выборочным данным

Оценка в статистике – это правило вычисления оцениваемого параметра. Она указывает приближенное значение показателей выборки относительно этих параметров генеральной совокупности. По мере увеличения числа наблюдений выборочные средние и другие параметры все больше приближаются к этим значениям генеральной совокупности.

Степень соответствия показателей оценивается *ошибкой* *(m)*. Ее запись производится вместе с оцениваемым параметром, например, $M \pm m_M$, $\sigma \pm m_\sigma$, $V \pm m_V$.

Ошибка указывает интервал, в пределах которого находится этот показатель в генеральной совокупности. Чем меньше ошибка, тем ближе значение выборочного показателя к этому показателю генеральной совокупности. Чем больше число наблюдений и чем однороднее выборка, тем меньшая ошибка среднего и других показателей.

1.4. Оценка статистических параметров по выборочным данным

Представление средней арифметической выборки приводится обязательно с ее ошибкой. Стандартная ошибка средней рассчитывается:

$$m_M = \sqrt{\frac{\sum (x_i - M_x)^2}{N(N-1)}}, \text{ или } m_M = \sqrt{\frac{\sigma^2}{N}}, \text{ или } m_M = \frac{\sigma}{\sqrt{N}} \qquad (1.12)$$

Ошибка среднеквадратического отклонения определяется по формуле:

$$m_\sigma = \sigma / \sqrt{2(N-1)}. \qquad (1.13)$$

Ошибка дисперсии вычисляется путем возведения в квадрат ошибки среднеквадратической.

Ошибка коэффициента вариации рассчитывается следующим образом:

$$m_V = \frac{V}{\sqrt{N}} \cdot \sqrt{\frac{1}{2} + (V/100)^2}. \qquad (1.14)$$

1.4. Оценка статистических параметров по выборочным данным

Поскольку параметр *m* характеризует ошибку утверждения (прогноза) о том, что выборочное среднее равно генеральному среднему, то чем выше требование к вероятности этого вывода, тем шире должен быть обеспечивающий точность такого прогноза интервал, называемый ***доверительным интервалом***. Его величина задается вероятностью безошибочного прогноза, которую принято *называть доверительной вероятностью* (*уровень вероятности,* надежность опыта, вероятность безошибочного прогноза). В исследованиях допускается доверительная вероятность (*P*) не менее 95 % (0,95 частей от 1). В этих случаях *P* для средних арифметических при достаточно большом числе наблюдений ($N > 30$) равен $\pm 2m$. Предельная ошибка выборки $\Delta = M \pm 2m$. При доверительной вероятности 99 % (0,99) доверительный интервал составит $\pm 3m$, $\Delta = M \pm 3m$. По иному, в отношении доверительного интервала можно сказать так: ***он показывает какой процент вариант выборки (выборок) подтверждает искомую статистическую закономерность***.

1.4. Оценка статистических параметров по выборочным данным

Каждому значению доверительной вероятности соответствует свой ***уровень значимости*** **(α)**. Он выражает вероятность нулевой гипотезы: вероятность того, что выборочная и генеральная средние не отличаются друг от друга. Иначе говоря, чем выше уровень значимости, тем меньше можно доверять утверждению, что различия существуют, т. е., ***он показывает, какой процент вариант совокупности (выборок) отвергают искомую статистическую закономерность***. Уровень значимости 5 % (0,05) дополняет доверительную вероятность 95 % (0,95). В сумме они составляют 100 % (1). Если доказано подобие между выборками при α = 5 % (0,05), то из этого следует, что до 5 % вариант выборки подобие не подтверждают. В таблицах приложения приводятся численные значения для *P* или α соответственно 0,95 и 0,99; 0,05 и 0,01. В этих случаях при интерпретации мы можем утверждать нулевую гипотезу (H_0). При более высоких уровне вероятности 0,99 и уровне значимости 0,01 мы получаем сильный довод для утверждения нулевой гипотезы.

1.4. Оценка статистических параметров по выборочным данным

Проверка статистических гипотез. Методологической основой любого исследования является формулировка рабочей гипотезы. В ходе исследования рабочая гипотеза либо принимается, либо отвергается. Статистической называют гипотезу о виде неизвестного распределения или о параметре распределения. Примеры гипотез:

- генеральная совокупность распределяется по закону Пуассона;
- средние арифметические двух совокупностей не равны между собой;
- дисперсии двух совокупностей равны между собой.

Выдвинутую гипотезу называют ***основной*** *или* ***нулевой*** (H_0). Гипотезу, которая противоречит нулевой, называют ***конкурирующей*** *или* ***альтернативной*** (H_1). Если нулевая гипотеза предполагает, что M = 20, то логическим отрицанием будет $M \neq 15$. Простая гипотеза содержит одно предположение, сложная – состоит из конечного или бесконечного множества простых гипотез.

1.4. Оценка статистических параметров по выборочным данным

Выдвинутую гипотезу проверяют на правильность ее статистическими методами, т. е. проводят статистическую проверку. При проверке могут быть допущены ошибки двух родов.

Ошибка первого рода – отвергается правильная гипотеза. Вероятность совершить ошибку первого рода называют *уровнем значимости* (α). Это значит, что в 5 случаях из 100 мы рискуем допустить ошибку первого рода.

Ошибка второго рода – принимается неправильная гипотеза, вероятность ошибки второго рода не имеет какого-то особого общепринятого названия, может обозначаться как β. Однако с этой величиной тесно связана другая, имеющая большое статистическое значение – мощность критерия. Таким образом, чем выше мощность, тем меньше вероятность совершить ошибку второго рода.

1.4. Оценка статистических параметров по выборочным данным

При проведении оценки обычно приходится идти на компромисс между приемлемым уровнем ошибок первого и второго рода. Зачастую для принятия решения используется пороговое значение, которое может варьироваться с целью сделать статистический тест более строгим или, наоборот, более мягким. Этим пороговым значением является уровень значимости, которым задаются при проверке статистических гипотез. Например, повышение чувствительности прибора приведёт к увеличению риска ошибки первого рода (ложное определение явления или *ложноположительный результат*), а понижение чувствительности – к увеличению риска ошибки второго рода (пропуск явления или *ложноотрицательный результат*).

1.4. Оценка статистических параметров по выборочным данным

Для проверки нулевых гипотез используют статистические критерии.

При сравнении дисперсий используют критерий Фишера.

В большинстве исследований для статистической проверки гипотез существенности различий средних арифметических используют параметрический критерий Стьюдента. Если нулевая гипотеза принимается, это не означает ее доказательство. Доказать на основании однократной или косвенной проверки гипотезу нельзя, а опровергнуть можно. Для повышения точности статистических данных необходимо уменьшить вероятности ошибок первого и второго рода, увеличить объем выборок. Область применения того или иного критерия задается законом его распределения.

1.4. Оценка статистических параметров по выборочным данным

Оценка точности опыта. При исследованиях *методического* характера необходимо приводить их оценку по показателю *точность опыта* (*p*). Его смысл состоит в установлении величины ошибки среднего арифметического (m_M) в процентах от величины среднего арифметического (M). Показатель точности опыта можно определить по одной из двух формул:

$$p = (m_M / M) \cdot 100; \quad p = V / \sqrt{N}, \qquad (1.15)$$

где V – коэффициент вариации.

1.4. Оценка статистических параметров по выборочным данным

Опыт считается достаточно точным, если p < 3 %, удовлетворительным – при его величине 3–5 %. Если величина точности опыта более 5 %, к полученным выводам следует относиться осторожно и увеличить число повторностей в опыте. Эти градации обязательны для полевых опытов с растениями. Некоторые приборы для анализа могут давать значительно большую погрешность (p до 15 %).

Ошибка показателя точности опыта вычисляется следующим образом:

$$m_p = \pm\, p\, \sqrt{(1/2N) + (p/100)^2} \qquad (1.16)$$