

Лекция 14

1. Основные задачи математической статистики
2. Основные понятия математической статистики
3. Статистические числовые характеристики
4. Статистический анализ параметров распределения
 - 4.1. Статистический анализ выборочного среднего
 - 4.2 Статистический анализ выборочной дисперсии
5. Доверительный интервал и доверительная вероятность
 - 5.1. Приближённая оценка доверительного интервала

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика – это наука, которая занимается сбором, обработкой и описанием экспериментальных данных. При этом ставится цель выявить закономерности в случайном явлении по результатам испытаний (наблюдений), зафиксированных в виде числовых данных. МС - является источником и базой теории вероятностей.

1. Основные задачи математической статистики

1. Определение на основе данных опыта (неизвестного) закона распределения случайной величины.
1. Определение по данным опыта неизвестных параметров распределения.
1. Проверка статистических гипотез.

Эффективность и надежность расчетов с использованием теории вероятностей определяется исходным экспериментальным материалом и, следовательно, не в последнюю очередь его организацией, то есть методами сбора и представления данных.

Например, для изучения геофизических явлений составляется международный геофизический график-календарь, который так регламентирует периоды наблюдения и их продолжительность, чтобы при минимальных экономических затратах в результатах наблюдений содержалась бы информация как о быстрых, так и медленных изменениях, например, в структуре атмосферы.

При контроле выпускаемых в больших количествах элементов электроники регламентируется порядок выбора и количество испытываемых экземпляров.

2. Основные понятия математической статистики

Генеральная совокупность

Предположим, исследуется некий признак. Например, параметр h_{11} в партии выпущенных фирмой транзисторов. Пусть в партии имеется N транзисторов, и все без исключения транзисторы подвергаются испытаниям. Тогда совокупность

всех без исключения N результатов испытаний называется генеральной совокупностью. Число N называется объемом генеральной совокупности. Только в этом случае мы получим исчерпывающую информацию.

Выборочная совокупность

На практике сплошных обследований обычно не делают. Чаще всего **выбирают** какую-то часть (или несколько частей). Отсюда появляется термин "выборочная совокупность". Выборочная совокупность с объемом выборки n ($n < N$), дает информацию об исследуемом признаке с некоторой погрешностью. Результат тем точнее, чем меньше n отличается от N . Такой метод обследования называется **выборочным**, и его обоснованием служит теорема Чебышёва.

Простой статистический ряд

Представляет собой первичные статистические данные в виде таблицы:

№	1	2	3	...	n
x_i	x_1	x_2	x_3	...	x_n

Примерами простого статистического ряда могут служить таблицы 1 и 3 расчетно-графического задания. Простой статистический ряд, если n достаточно велико, является громоздким и неудобным для обработки.

Статистическая совокупность

Если простой статистический ряд достаточно большой (например, около 1000 значений), его преобразуют в статистическую совокупность. Для этого всю область значений, в пределах которой лежат результаты испытаний, разбивают на k интервалов (разрядов, групп). Разбиение делают произвольно, преследуя цель упростить обработку. Затем подсчитываются частоты, с которыми случайная величина попадает в выбранные интервалы.

Таблица 1

Интервал	$x_1 - x_2$	$x_2 - x_3$	...	$x_k - x_{k+1}$
Частота	n_1	n_2	...	n_k
Относительная частота	n_1/n	n_2/n	...	n_k/n

Пример такого разбиения представляет таблица 4 расчетно-графического задания.

3. Статистические числовые характеристики

Числовые характеристики, полученные при обработке результатов эксперимента, называются **статистическими числовыми характеристиками**, или **статистическими оценками**, или просто **оценками**

Числовая характеристика	(Статистическая) оценка
m	m^*
D	D^*
α_k	α_k^*
μ_k	μ_k^*

m^* – среднее арифметическое (выборочное среднее);

D^* – выборочная дисперсия;

α_k^* – оценка начального момента k -го порядка;

μ_k^* – оценка центрального момента k -го порядка.

$m^* = \frac{1}{n} \sum_{i=1}^n x_i = \bar{m}$; x_i – значения, взятые из простого статистического ряда;

$\bar{m} = \bar{x}_1 p_1^* + \bar{x}_2 p_2^* + \dots + \bar{x}_k p_k^*$; – результаты, взятые из статистической совокупности $\bar{x}_i = \frac{x_i + x_{i+1}}{2}$;

$D^* = \frac{1}{n} \sum_{i=1}^n (x_i - m^*)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (m^*)^2$ x_i – результаты, взятые из простого статистического ряда.

4. Статистический анализ параметров распределения

Полученные при обработке выборочной совокупности с объемом n ($n < \infty$) оценки числовых характеристик случайных величин являются случайными числами, поэтому их подвергают статистическому анализу по трём направлениям:

- устанавливают отсутствие или наличие **смещенности**;
- оценивают **состоятельность** полученных оценок;
- оценивают их **эффективность**.

Несмещенной называют статистическую оценку, когда её математическое ожидание равно оцениваемому параметру при любом объёме выборки

$$M[U^*] = U.$$

Смещенной называют оценку, если её математическое ожидание не равно оцениваемому параметру.

Следует отметить, что даже в отсутствие систематической ошибки вследствие большого разброса данных полученная оценка может сильно не совпадать с оцениваемой числовой характеристикой.

Оценка называется **состоятельной**, если при увеличении объёма выборки она стремится к оцениваемому параметру.

Эффективной называют статистическую оценку, которая при заданном объёме выборки имеет наименьшую возможную дисперсию.

4.1. Статистический анализ выборочного среднего

Пусть измеряется случайная величина X . Сделаны замеры $x_1, x_2, \dots, x_n$. Найдено выборочное среднее

$$m^* = \frac{1}{n} (x_1 + x_2 + \dots + x_n).$$

Представим, что $x_1, x_2, \dots, x_n$ — это конкретные реализации функции X , и таким образом математическое ожидание нашей оценки

$$M[m^*] = M\left[\frac{1}{n} (x_1 + x_2 + \dots + x_n)\right] = \frac{n \cdot m}{n} = m.$$

Как и следовало ожидать, **выборочное среднее является несмещенной оценкой** математического ожидания

Состоятельность. Для её проверки используем теорему Чебышева

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n x_i - m_x\right| \leq \varepsilon\right) \xrightarrow{n \rightarrow \infty} 1.$$

Эффективность или неэффективность оценки математического ожидания зависит от вида распределения величины X . Если закон распределения нормальный, то оценка математического ожидания имеет наименьшую дисперсию. Есть такие законы распределения, когда это не так.

Дисперсия полученной оценки выборочного среднего (что одно и то же — среднее арифметическое) как было ранее найдено (см. доказательство теоремы Чебышёва)

$$D[m^*] = D/n$$

4.2 Статистический анализ выборочной дисперсии

По определению выборочная дисперсия рассчитывается согласно выражениям

$$D^* = \frac{1}{n} \sum_{i=1}^n (x_i - m^*)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (m^*)^2.$$

Оценка выборочной дисперсии является **состоятельной**, так как первое слагаемое при увеличении **объема выборки n** стремится к математическому ожиданию квадрата, а второе - к квадрату математического ожидания, следовательно, D^* стремится по вероятности к D :

$$\lim_{n \rightarrow \infty} M D^* = M[X^2] - m^2; \quad \lim_{n \rightarrow \infty} M D^* = D$$

Оценка выборочной дисперсии является **смещённой**:

$$D^* = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left[\frac{1}{n} (x_1 + x_2 + \dots + x_n) \right]^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{1}{n^2} \sum_{i=1}^n x_i^2 - 2 \sum_{i \neq j} x_i x_j.$$

Математическое ожидание выборочной дисперсии D^* найдём, используя свойство дисперсии:

$$D(X + C) = D(X).$$

Для этого в полученном выражении из каждого x_i вычтем математическое ожидание m , т.е. используем подстановку $C = -m$. Принимая во внимание, что $M(X - m) = 0$, получим:

$$M D^* = \left(\frac{1}{n} - \frac{1}{n^2} \right) M \left[\sum_{i=1}^n (x_i - m)^2 \right] - 2M \left[\sum_{i \neq j} (x_i - m)(x_j - m) \right]; \quad M D^* = \frac{n-1}{n} D.$$

Оценка выборочной дисперсии оказывается смещенной, и, как следует из полученного выражения, она занижена. Если **n не велико**, то используют т.н. **исправленную оценку**: $\tilde{D} = \frac{n}{n-1} D^*.$

5. Доверительный интервал и доверительная вероятность

Полученные оценки на числовой оси отображаются точкой, и потому их принято называть точечными. Оценки могут быть достаточно случайными, и они не дают ответа на вопрос, как далеки они могут быть от истинных значений числовой характеристики. Т.е. требуется знать - к каким ошибкам может привести замена неизвестного истинного параметра α его точечной оценкой α^* .

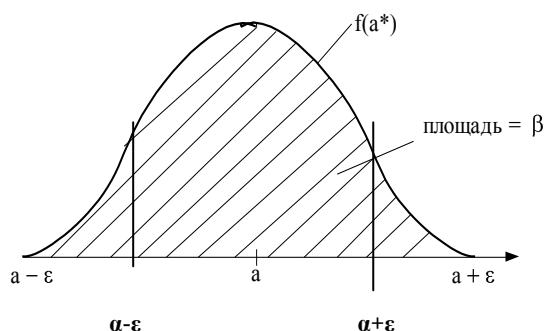
Чтобы получить представление о точности и надежности полученной оценки в математической статистике пользуются взаимосвязанными понятиями **доверительного интервала и доверительной вероятности**.

Пусть в статистических расчетах параметра α получена его оценка α^* . Назначается некоторая достаточно высокая вероятность β того, чтобы считать практически достоверным событие "разница между полученной оценкой и истинным значением параметра меньше ε ", а именно:

$$P(|\alpha^* - \alpha| \leq \varepsilon) = \beta, \quad (\beta = 0.8 - 0.95)$$

Допустим, что нам известны математическое ожидание (т.е. истинное значение) параметра α и дифференциальная функция распределения оценки α^* . Тогда связь между значениями β и ε определяется выражением:

$$\beta = \int_{\alpha - \varepsilon}^{\alpha + \varepsilon} f(\alpha^*) d\alpha^* = F(\alpha + \varepsilon) - F(\alpha - \varepsilon) \quad (**)$$



Геометрическое толкование выражения заключается в том, что оно связывает β – вероятность появления α^* в интервале $[\alpha - \varepsilon, \alpha + \varepsilon]$ – с его границами $\alpha - \varepsilon$ и $\alpha + \varepsilon$. Говорят, что "с вероятностью β значение параметра попадает в интервал $[\alpha - \varepsilon, \alpha + \varepsilon]$ ".

Доверительный интервал можно рассматривать как интервал значений искомого параметра не противоречащего опытными данным. В качестве примера можно привести правило "трех сигм". С доверительной вероятностью 0,997 нормально распределенная случайная величина не выходит за пределы шести сигм, с вероятностью 0,95 такая случайная величина не выходит за пределы четырех сигм. На практике последнее означает, что в интервале значений $\pm 4\sigma$ должно находиться около 95% представленного экспериментального материала.

К сожалению, выражение β не может быть использовано на практике, т.к. значение α заранее неизвестно. Поэтому полученное значение α^* принимают как истинное и вводят следующие обозначения

β – **доверительная вероятность** ;

$I_\beta = (\alpha^* - \varepsilon, \alpha^* + \varepsilon)$ – **доверительный интервал**;

$\alpha - \varepsilon, \alpha + \varepsilon$ – **границы доверительного интервала**

Например: $I_{0,95} = 7,5 - 0,02, 7,5 + 0,02 = 7,5 \pm 0,02$.

Говорят: "С вероятностью β истинное значение параметра α попадает в интервал $(\alpha^* - \varepsilon, \alpha^* + \varepsilon)$ " или "с вероятностью 0,95 истинное значение параметра α лежит в пределах $7,5 \pm 0,02$ ".

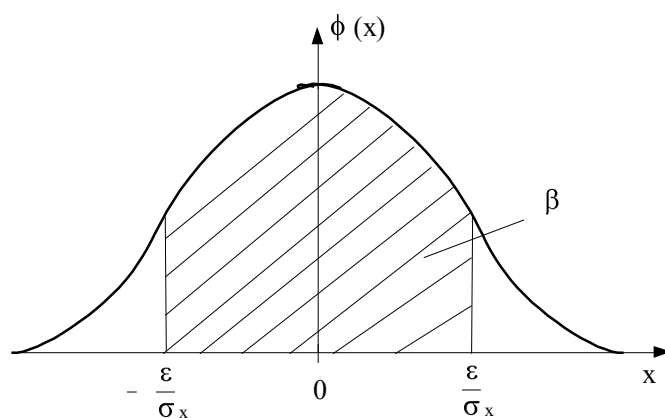
5.1. Приближённая оценка доверительного интервала

При приближённых оценках доверительных границ принимают:

- найденное значение α^* соответствует истинному значению α ;
- при неизвестном законе распределения случайной величины α^* его

принимают как нормальный.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$



Ноль координат соответствует оценке параметра α^* .

$$\beta = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz = 2\Phi(x); \quad x = \frac{\varepsilon}{\sigma_x}.$$

β	0,8	0,9	0,95	0,97	0,99	0,997
x	1,282	1,643	1,96	2,169	2,58	3

$$\varepsilon = x \cdot \sigma_x.$$