Слайд 1. 5.2. Введение в математическую статистику

5.2.1. Предмет и задачи математической статистики

Предметом математической статистики является разработка методов регистрации, описания и анализа статистических экспериментальных данных, получаемых в результате наблюдения массовых случайных явлений.

Математическая статистика как наука основывается на теории вероятностей. И теория вероятностей, и математическая статистика занимаются количественным и качественным анализом случайных явлений. Существенным же различием между ними является то, что теория вероятностей исходит из предположения, что закон распределения случайной величины известен, известны вероятности наступления отдельных событий, по которым можно рассчитать вероятности некоторых других событий. А зная математическую модель явления, можно предсказать возможное течение явления.

Задачи математической статистики являются, в известной мере, обратными к задачам теории вероятностей. В математической статистике исходят из известных реализаций каких-либо случайных событий, из так называемых статистических данных, и на основании этих данных строят математическую модель, т.е. устанавливают вероятности событий и законы распределения случайных величин.

Перечислим основные задачи математической статистики, которые наиболее важны по своим практическим применениям:

1. Выбор способов сбора и группировки статистических сведений.
2. Определение законов распределения случайной величины по статистическим данным.
3. Проверка правдоподобия гипотез.
4. Нахождение неизвестных параметров распределения.

Все задачи математической статистики касаются вопросов обработки наблюдений над массовыми случайными явлениями, но в зависимости от характера решаемого практического вопроса и от объема имеющегося экспериментального материала эти задачи принимают ту или иную форму реализации.

Теоретически при достаточном количестве опытов свойственные изучаемым случайным величинам закономерности будут проявляться достаточно четко. Но на практике часто приходится иметь дело с ограниченным количеством экспериментальных данных; в связи с этим результаты наблюдений и их обработка всегда содержат элемент случайности. Поэтому возникает вопрос, какие черты наблюдаемого явления относятся к

постоянным, устойчивым и действительно присущи ему, а какие являются случайными и проявляются в данной серии наблюдений только за счет ограниченного объема экспериментальных данных. Естественно, что методика обработки экспериментальных данных, по возможности, должна сохранять типичные, характерные черты наблюдаемого явления и отбрасывать все несущественное, связанное с недостаточным объемом опытного материала. Возникает характерная для математической статистики задача сглаживания или выравнивания статистических данных, представления их в наиболее компактном виде с помощью простых аналитических зависимостей.

Задача проверки правдоподобия гипотез тесно связана с предыдущей. Основной вопрос, который может возникнуть: согласуются ли результаты эксперимента с гипотезой о том, что данная случайная величина подчинена тому или иному закону распределения, т.е. вопрос согласованности теоретического и статистического распределения. Как бы хорошо ни было подобрано теоретическое распределение, между ним и статистическим распределением неизбежны некоторые расхождения. Естественно возникает другой вопрос: объясняются ли эти расхождения только случайными обстоятельствами, связанными с ограниченным числом наблюдений, или они являются существенными и связаны с тем, что подобранное нами теоретическое распределение плохо выравнивает данное статистическое распределение. Для решения подобных вопросов математическая статистика выработала ряд специальных приемов.

Часто при обработке экспериментального материала характер закона распределения известен до опыта, из теоретических соображений. Тогда возникает задача определения некоторых параметров (числовых характеристик) случайной величины или системы случайных величин. При небольшом числе опытов экспериментальный материал неизбежно содержит в себе элемент случайности; поэтому случайными оказываются и все параметры, вычисленные на основе этих данных. В таких условиях может быть поставлена только задача об определении оценок или подходящих значений для искомых параметров, т.е. таких приближенных значений, которые бы в среднем приводили к меньшим ошибкам, чем всякие другие. Поэтому с задачей отыскания подходящих значений числовых характеристик тесно связана задача оценки их точности и надежности.

5.2.2. Генеральная совокупность

Пусть требуется изучить совокупность однородных объектов относительно некоторого качественного или количественного признака, характеризующего

эти объекты. Например, при изучении партий деталей качественным признаком может служить стандартность деталей, а количественным – контролируемый размер детали. С этой целью проводят сплошное обследование объектов совокупности относительно интересующего нас признака.

Слайд 2. Определение. Вся подлежащая изучению совокупность объектов называется *генеральной совокупностью*. Число элементов в генеральной совокупности называется ее объемом.

Однако, сплошное наблюдение, т.е. изучение всех членов совокупности, не является единственно возможным способом получения о ней достаточно точной информации и в ряде случаев нецелесообразным. Кроме того, часто экономически невыгодно производить обследование всей совокупности, если по результатам изучения сравнительно небольшой ее части можно получить с достаточной для практики достоверностью необходимую информацию о всей совокупности. В таких случаях отбирают из всей совокупности ограниченное число объектов и подвергают их изучению. Такой метод исследования называется *выборочным*, а отобранная совокупность объектов – выборкой.

Чтобы иметь право судить о генеральной совокупности по выборке, последняя должна быть образована случайно.

Охарактеризуем ошибки, которые могут возникнуть в ходе исследования.

Ошибкой регистрации называется разность между истинным и наблюденным значениями изучаемого признака у членов совокупности.

Ошибки регистрации могут быть систематическими и случайными. Систематические ошибки регистрации возникают при умышленном или неумышленном искажении изучаемого признака у членов совокупности в одну и ту же сторону (завышения или занижения).

Ошибкой репрезентативности называется расхождение характеристик признака в генеральной и выборочной совокупностях, возникающее только в результате того, что исследуется не вся совокупность, а лишь ее часть.

Систематическая ошибка репрезентативности возникает при нарушении случайности отбора членов в выборочную совокупность (например, если в выборку, образованную с целью прогноза урожайности, будут включены лишь наиболее урожайные или, наоборот, наименее урожайные участки).

Слайд 3. 5.2.3. Вариационные ряды и их характеристики

Предположим, что изучается некоторая случайная величина X, закон распределения которой в точности не известен, и требуется определить этот закон из опыта или проверить экспериментально гипотезу о том, что величина X подчинена тому или иному закону. С этой целью над случайной величиной X

производят ряд независимых опытов (наблюдений), в каждом из которых случайная величина X принимает определенное значение. Совокупность наблюдаемых значений $x_1, x_2, ..., x_n, ...$ представляет собой первичный статистический материал, подлежащий обработке и анализу. Например, регистрация числа баллов, полученных на вступительных экзаменах по математике 60 абитуриентами, дала следующие результаты:

20	19	22	24	21	18	23	17	20	16	15	23
21	24	21	18	23	21	19	20	24	21	20	18
17	22	20	16	22	18	20	17	21	17	19	20
20	21	18	22	23	21	25	22	20	19	21	24
23	21	19	22	21	19	20	23	22	25	21	21.

Как видим, у абитуриентов неодинаковое число баллов, хотя много совпадающих. Для дальнейшей целенаправленной обработки первичного статистического материала важно знать, как часто встречаются различные значения признака в генеральной совокупности.

Слайд 4. Введем ряд понятий.

— Различные значения признака x_i, наблюдающиеся у членов совокупности, называются *вариантами*.

— Число, показывающее, сколько раз встречается вариант в совокупности, называется его *частотой* (m_i).

— Отношение частоты варианта m_i к числу членов совокупности n называется его *частостью*: $p_i^* = \frac{m_i}{n}$. Частости вариантов выражают доли (удельные веса) членов совокупности с одинаковыми значениями признака.

— Частоты или частости вариантов называют их *весами*.

Выделить все наблюдаемые варианты и подсчитать соответствующие им частоты можно с помощью, так называемой, рабочей таблицы. В первом столбце указывают варианты в порядке их возрастания или убывания, во второй последовательно заносят наблюдающиеся значения признака. Заполненная рабочая таблица в рассмотренном примере выглядит так:

Таблица 1

Число баллов (варианты)	Число абитуриентов (частоты)	Доля абитуриентов	
		Частость	В процентах
15	1	0,017	1,7
16	2	0,033	3,3
17	4	0,067	6,7
18	5	0,083	8,3
19	6	0,1	10,0

20	10	0,167	16,7
21	13	0,217	21,7
22	7	0,116	11,6
23	6	0,1	10,0
24	4	0,167	6,7
25	2	0,033	3,3
Итого	**60**	**1,00**	**100,0**

Слайд 5. Сумма частот дает объем совокупности (число абитуриентов – 60).

Определение. Вариационным рядом называется ранжированный в порядке возрастания или убывания ряд вариантов с соответствующими им весами.

Таким образом, в таблице 1 приведен вариационный ряд распределения 60 абитуриентов по числу баллов, полученных ими на приемных экзаменах.

В зависимости от того, какие значения может принимать признак, вариационные ряды делятся на дискретные (или прерывные) и непрерывные (или интервальные).

Вариационный ряд называется *дискретным*, если значения признака отличаются друг от друга не менее, чем на некоторую постоянную величину, и *непрерывным*, если значения признака могут отличаться на сколь угодно малую величину и непрерывно заполняют некоторый промежуток.

Вариационный ряд, приведенный в таблице 1, является дискретным. В общем виде дискретный вариационный ряд представляется таблицей 2.

Слайд 6.

Таблица 2

Варианты	Веса (частоты или частости)
x_1	m_1
x_2	m_2
…	…
x_m	m_m
Итого	n

Примерами непрерывных вариационных рядов могут служить: распределение рабочих предприятия по проценту выполнения нормы, людей по возрасту, посевной площади по урожайности и т.п. Непрерывный вариационный ряд задается таблицей 3.

Таблица 3

Значения признака	Веса (частоты или частости)
От x_1 *до* x_2	m_1
От x_2 *до* x_3	m_2
	…

… *От* x_m *до* x_{m+1}	m_m
Итого	n

Слайд 7. Предполагается, что каждому интервалу принадлежит лишь один из его концов – либо во всех случаях левый, либо во всех случаях правый. Будем считать для определенности, что в таблице 2 варианты расположены в возрастающем порядке, а веса в таблицах 2 и 3 отличны от нуля.

Разности $x_2 - x_1,\ x_3 - x_2, \ldots, x_{m+1} - x_m$ называются *интервальными разностями*.

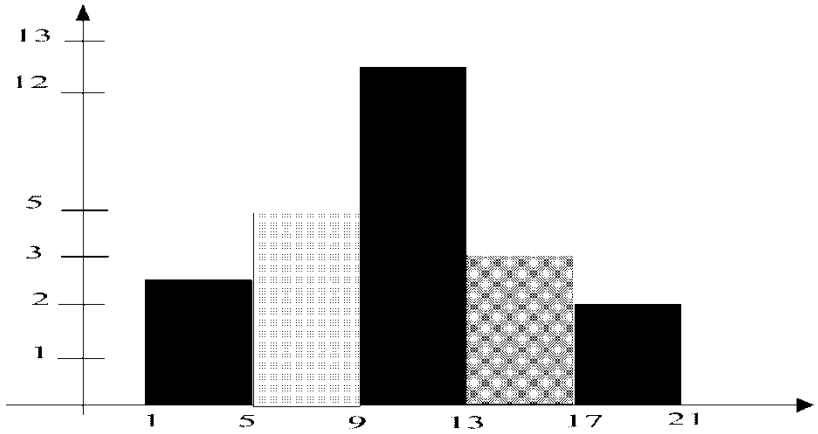
Вариационный ряд часто оформляется графически в виде гистограммы. По оси абсцисс откладываются варианты, и на каждой из вариант как на основании строится прямоугольник, площадь которого равна частоте данного варианта. В качестве высоты прямоугольника берется частное частоты каждого варианта на его длину. Из способа построения гистограммы следует, что полная ее площадь равна единице.

Пример. Построим гистограмму частот по данному распределению выборки.

номер интервала	частичный интервал	сумма частот вариант интервала	плотность частоты
i	$x_i; x_{i+1}$	m_i	$\frac{m_i}{h_i}$
1	1 – 5	10	2,5
2	5 – 9	20	5
3	9 – 13	50	12,5
4	13 – 17	12	3
5	17 – 21	8	2

Слайд 8. Р е ш е н и е: $h_i = x_{i+1} - x_i; n = \sum_{i=1}^{5} m_i = 100$.

Построим на оси абсцисс заданные интервальные длины $h_i = 4$. Проведем над этими интервалами отрезки, параллельные оси абсцисс, и находящиеся от нее на расстояниях, равных соответствующим $\frac{m_i}{h_i}$. Искомая гистограмма имеет вид:



Замечание. В теории вероятностей под распределением понимают соответствие между возможными значениями случайной величины и их вероятностями, а в математической статистике – соответствие между наблюдаемыми вариантами и весами (частотами или частостями).

Слайд 9. Очевидно, что при увеличении числа опытов можно выбирать более мелкие разряды; при этом гистограмма будет все более приближаться к некоторой кривой, ограничивающей площадь, равную единице. Нетрудно убедиться, что эта кривая представляет собой график плотности распределения величины X.

5.2.4. Эмпирическая функция распределения

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$:

$$F^*(x) = \frac{m(x)}{n},$$

где $m(x)$ – число вариант, меньших x; n – объем выборки.

Эмпирическая функция обладает следующими *свойствами*:

1. Значения эмпирической функции принадлежат отрезку [0;1];
2. $F^*(x)$ – неубывающая функция;
3. Если x_1 – наименьшая варианта, а x_k – наибольшая, то $F^*(x) = 0$ при $x < x_1$ и $F^*(x) = 1$ при $x > x_k$.

Слайд 10. В теории вероятностей под функцией распределения случайной величины X понимают вероятность того, что в результате опыта случайная величина X примет значение меньше x. Число же $F^*(x)$ есть относительная частота этого события. По теореме Бернулли при больших n относительная частота события А приблизительно равна его вероятности (точнее, сходится по

вероятности к P(A)). Поэтому эмпирическая функция распределения $F^*(x)$ заменяет в математической статистике известную из теории вероятностей функцию распределения $F(x)$, называемую теоретической функцией распределения.

Пример. Найти эмпирическую функцию по данному распределению выборки.

x_i	1	4	6
m_i	10	15	25

Р е ш е н и е. Найдем объем выборки:

$$n = 10 + 15 + 25 = 50.$$

Наименьшая варианта равна единице, следовательно, $F^*(x) = 0$ при $x \le 1$.

Значение $x < 4$, а именно $x_1 = 1$, наблюдалось 10 раз, следовательно, $F^*(x) = 10/50 = 0.2$ при $1 < x \le 4$.

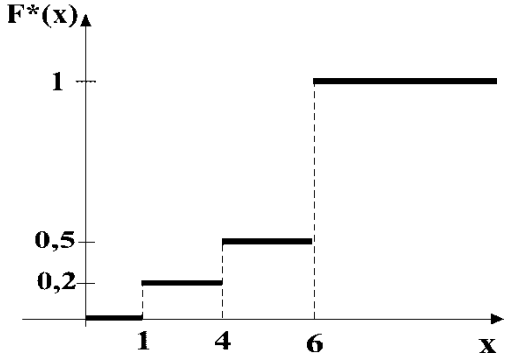
Слайд 11. Значение $x < 6$, а именно $x_1 = 1$ и $x_2 = 4$, наблюдалось 10+15=25 раз, следовательно, $F^*(x) = 25/50 = 0.5$ при $4 < x \le 6$.

x = 6 – наибольшая варианта, поэтому $F^*(x) = 1$ при $x > 6$.

Искомая эмпирическая функция имеет вид:

$$F^*(x) = \begin{cases} 0 & при \ \ x \le 1, \\ 0{,}2 & при \ \ 1 < x \le 4, \\ 0{,}5 & при \ \ 4 < x \le 6, \\ 1 & при \ \ x > 6. \end{cases}$$

График этой функции имеет вид:



Слайд 12. 5.2.5. Числовые характеристики вариационного ряда

Пусть собранный и обработанный статистический материал представлен в виде вариационного ряда. Далее необходимо подвергнуть его анализу.

Желательно охарактеризовать ряд с помощью некоторых постоянных, которые представляли бы его в целом и отражали присущие изучаемой совокупности закономерности. К таким постоянным относятся средние: средняя арифметическая, средняя геометрическая, средняя гармоническая, средняя квадратическая и др. Из них первой по значимости является средняя арифметическая.

Определение. *Средней арифметической* вариационного ряда называется дробь, числителем которой служит сумма произведений вариантов ряда на соответствующие им веса, а знаменателем – сумма весов, т.е.

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_m n_m}{n} = \frac{\sum_{i=1}^{m} x_i\, n_i}{n}. \qquad (1)$$

Пример. Вычислить среднюю заработную плату рабочих по данным таблицы.

Заработная плата, тыс. р.	Число рабочих	Заработная плата, тыс. р.	Число рабочих
70 – 80	1	110 – 120	20
80 – 90	3	120 – 130	12
90 – 100	10	130 – 140	7
100 - 110	15	140 – 150	2

По формуле (1) находим: **Слайд 13.**

$$\bar{x} = \frac{75 \cdot 1 + 85 \cdot 3 + 95 \cdot 10 + 105 \cdot 15 + 115 \cdot 20 + 125 \cdot 12 + 135 \cdot 7 + 145 \cdot 2}{70} \approx 79{,}86 \cdot$$

За среднюю арифметическую непрерывного вариационного ряда принимают среднюю арифметическую дискретного распределения, соответствующего данному непрерывному, т.е. частоты непрерывного распределения относят к серединам соответствующих интервалов, которые становятся вариантами.

Нахождение средней арифметической непрерывного вариационного ряда осложняется, если крайние интервалы распределения не замкнуты. В этом случае сначала устанавливают границы этих интервалов, считая, что интервальная разность первого интервала такая же, как у второго, а у последнего – такая же, как у предпоследнего.

Пример. Вычислить среднее число жителей в поселках городского типа СНГ по данным таблицы.

Число жителей, Тыс. человек	**Число поселков городского типа**

Менее 3	1039
От 3 до 5	976
5–10	1251
10–20	422
20 и более	51
Итого	**3739**

Установим границы крайних интервалов. Последнему интервалу предшествует интервал от 10 до 20 тыс. человек. Его интервальная разность равна 10 тыс. человек. **Слайд 14.** Следовательно, условно считаем правую границу последнего интервала равной 20+10=30.

Аналогично рассуждая, получим, что начало первого интервала равно 1. Учитывая эти результаты и принимая за варианты середины интервалов, найдем искомое среднее число жителей:

$$\bar{x} = \frac{2 \cdot 1039 + 4 \cdot 976 + 7{,}5 \cdot 1251 + 15 \cdot 422 + 25 \cdot 51}{3739} \approx 6{,}14 \text{ (тыс. человек)}$$

Сформулируем теоремы, характеризующие свойства средней арифметической.

Теорема 1. Если варианты увеличить (уменьшить) в одно и то же число раз, то средняя арифметическая увеличится (уменьшится) во столько же раз, т.е.

$$\overline{k \cdot x} = \frac{\sum_{i=1}^{m} k \cdot x_i \cdot n_i}{n} = k \cdot \bar{x}. \qquad (2)$$

Теорема 2. Если варианты уменьшить (увеличить) на одно и то же число, то средняя арифметическая уменьшится (увеличится) на то же число, т.е.

$$\frac{\sum_{i=1}^{m} (x_i - c) \cdot n_i}{n} = \bar{x} - c. \qquad (3)$$

Слайд 15. Теорема 3. Сумма произведений отклонений вариантов от средней арифметической на соответствующие им веса равна нулю.

$$\sum_{i=1}^{m} (x_i - \bar{x}) \cdot n_i = 0. \qquad (4)$$

Теорема 4. При увеличении или уменьшении весов в одно и то же число раз средняя арифметическая не изменяется

$$\frac{\sum_{i=1}^{m} x_i \cdot k \cdot n_i}{k \cdot n} = \bar{x}. \qquad (5)$$

Пусть некоторая совокупность разбита на части – группы, не обязательно одинаковые по объему. Тогда средние арифметические распределения членов групп называется *групповыми средними*, а среднюю арифметическую распределения по тому же признаку всей совокупности – *общей средней*.

Определение. Группы называются непересекающимися, если каждый член совокупности принадлежит только одной группе.

Теорема 5. Общая средняя равна средней арифметической групповых средних всех непересекающихся групп.

Слайд 16.

$$\bar{x} = \frac{\overline{x_1} \cdot N_1 + \overline{x_2} \cdot N_2 + \ldots + \overline{x_l} \cdot N_l}{N_1 + N_2 + \ldots + N_l} = \frac{\sum_{i=1}^{l} \overline{x_i} \cdot N_i}{N},$$

где $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_l$ – групповые средние; $N_1, N_2, \ldots, N_l$ – объем соответствующих непересекающихся групп.

Теорема 6. Если каждое значение признака z представляет сумму (разность) значений признаков x и y, то средняя арифметическая признака z равна сумме (разности) средних арифметических x и y.

Пусть $z_i = x_i \pm y_i \ (i = 1,2,\ldots,n)$. Тогда

$$\bar{z} = \frac{\sum_{i=1}^{m} z_i \cdot n_i}{n} = \frac{\sum_{i=1}^{m} (x_i \pm y_i) \cdot n_i}{n} = \frac{\sum_{i=1}^{m} x_i \cdot n_i}{n} \pm \frac{\sum_{i=1}^{m} y_i \cdot n_i}{n} = \bar{x} + \bar{y}. \qquad (6)$$

Слайд 17. Дисперсия вариационного ряда и ее свойства

Средние – это постоянные величины, которые определенным образом характеризуют распределение. О некоторых распределениях судят только по средним. Например, для сравнения уровней заработной платы в различных отраслях промышленности достаточно сравнить средние заработные платы в них. Однако по средним нельзя судить ни о различиях между уровнями заработной платы наиболее высоко и низкооплачиваемых работников, ни о том, какие отклонения от средней заработной платы имеют место.

В статистике наибольший интерес представляет разброс значений признака около их средней арифметической. Отклонения вариантов x_i от средней арифметической выражают разности $x_i - \bar{x}$, а веса вариантов показывают, как часто эти разности встречаются в распределении.

На практике и в теоретических исследованиях рассеяние признака чаще характеризуют дисперсией и средним квадратическим отклонением.

Определение. Дисперсией σ^2 вариационного ряда называется средняя арифметическая квадратов отклонений вариантов от их средней:

$$\sigma^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 \cdot n_i}{n}. \qquad (7)$$

Средним квадратическим отклонением называется арифметическое отклонение значение корня квадратного из дисперсии.

Пример. Вычислить дисперсию и среднее квадратическое отклонение распределения рабочих предприятия по времени, затрачиваемому на обработку одной детали.

Слайд 18.

Время, затрачиваемое на обработку одной детали, мин.	**Число рабочих**
2 – 4	42
4 – 6	73
6 – 8	154
8 – 10	205
10 – 12	26
Итого	**500**

Вычислим среднюю арифметическую, приняв за варианты середины интервалов:

$$\bar{x} = \frac{3 \cdot 42 + 5 \cdot 73 + 7 \cdot 154 + 9 \cdot 205 + 11 \cdot 26}{500} = 7{,}4 \text{ (мин.)}$$

По формуле (7) находим искомую дисперсию

$$\sigma^2 = \frac{(3-7{,}4)^2 \cdot 42 + (5-7{,}4)^2 \cdot 73 + (7-7{,}4)^2 \cdot 154 + (9-7{,}4)^2 \cdot 205 + (11-7{,}4)^2 \cdot 26}{500} = 4{,}24 \cdot$$

Среднее квадратическое отклонение того же распределения составляет

$$\sigma = \sqrt{4{,}24} \approx 2{,}059 \text{ (мин.)}.$$

Сравнив средние квадратические отклонения одного и того же признака в разных совокупностях, можно сказать, где вариация признака больше. Поэтому среднее квадратическое отклонение одновременно является и показателем однородности совокупности.

Слайд 19. Сформулируем теоремы, характеризующие свойства дисперсий. Предполагаем, что задан дискретный ряд, представленный таблицей 2, его средняя арифметическая равна $\bar{x}$, а дисперсия σ^2.

Теорема 1. Если все варианты увеличить (уменьшить) в k раз, то дисперсия увеличится (уменьшится) в k^2 раз, а среднее квадратическое отклонение – в $|k|$ раз.

$$\frac{\sum_{i=1}^{m}\left(k\cdot x_i - k\cdot\bar{x}\right)^2\cdot n_i}{n} = k^2\cdot\sigma^2.$$

Теорема 2. Если варианты увеличить или уменьшить на одну и ту же постоянную величину, то дисперсия не изменится.

$$\frac{\sum_{i=1}^{m}\left[\left(x_i + c\right) - \left(\bar{x} + c\right)\right]^2\cdot n_i}{n} = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2\cdot n_i}{n}.$$

Теорема 3. Если веса увеличить или уменьшить в одно и то же число раз, то дисперсия не изменится.

Слайд 20.

$$\frac{\sum\left(x_i - \bar{x}\right)^2\cdot k\cdot n_i}{\sum_{i=1}^{m} k\cdot n_i} = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2\cdot n_i}{n}.$$

Теорема 4. Дисперсия равна средней арифметической квадратов вариантов на соответствующие им веса без квадрата средней арифметической, т.е.

$$\sigma^2 = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n} - \left(\bar{x}\right)^2. \qquad (5.8)$$

Применение этой формулы упрощает вычисление дисперсии, если варианты невелики.

Пусть совокупность разбита на l непересекающихся групп.

Определение. Групповой дисперсией σ_j^2 называется дисперсия распределения членов j-ой группы относительно их средней – групповой средней $\bar{x}_j$, т.е.

$$\sigma_j^2 = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}_j\right)^2 m_i}{N_j}, \qquad (5.9)$$

где m_i – частоты вариантов в группе, $N_j = \sum_{i=1}^{m} m_i$ – объем группы.

Слайд 21. Определение. Дисперсия распределения по этому же признаку всей совокупности относительно общей средней называется *общей дисперсией*.

Определение. *Межгрупповой дисперсией* σ^2 называется средняя арифметическая квадратов отклонений групповых средних $\bar{x}_j$ всех непересекающихся групп от общей средней $\bar{x}$, т.е.

$$\delta^2 = \frac{\sum_{i=1}^{l} (\bar{x}_j - \bar{x})^2 \cdot N_j}{n}, \qquad (10)$$

где $N_j \ (j = 1,2,\ldots,l)$ – объемы групп.

Определение. Средней групповых дисперсий $\overline{\sigma^2}$ называется средняя арифметическая групповых дисперсий, т.е.

$$\overline{\sigma^2} = \frac{\sum_{j=1}^{l} \sigma_j^2 \cdot N_j}{n},$$

где $N_j \ (j = 1,2,\ldots,l)$ – объем непересекающихся групп.

Слайд 22. Теорема 5. (правило сложения дисперсий). Общая дисперсия σ^2 равна сумме средней групповых дисперсий $\overline{\sigma^2}$ непересекающихся групп, на которые разбита совокупность, и межгрупповой дисперсии δ^2, т.е.

$$\sigma^2 = \overline{\sigma^2} + \delta^2. \qquad (11)$$

Отношение среднего квадратичного отклонения к средней величине признака, вычисленное в процентах, называется *коэффициентом вариации*:

$$v = \frac{\sigma}{\bar{x}} \cdot 100\%. \qquad (12)$$

Разность между наибольшим и наименьшим значениями признака называют *размахом вариации*:

$$R = x_{\max} - x_{\min}.$$

Конец Части 1

Слайд 23. 5.2. Введение в математическую статистику. Часть 2

Моменты вариационного ряда

Моментом k–го порядка $M_k(a)$ варьирующего признака X по отношению к значению a называют среднее математическое из k-х степеней отклонений значений признака от a, т. е.

$$M_k(a) = \overline{(X-a)^k} = M(X-a)^k .$$

Если $a = 0$, момент называется *начальным* ν_k, а при $a = \overline{X}$ его называют *центральным* μ_k. Таким образом,

$$\nu_k = M(X^k) = \sum_{i=1}^{l} x_i^k \cdot \frac{m_i}{n},$$

$$\mu_k = M(X - \overline{X})^k = \sum_{i=1}^{l} (x_i - \overline{X})^k \cdot \frac{m_i}{n}.$$

За показатель отклонения распределения признака X от симметрии относительно X принимают величину

$$\alpha = \mu_3 / \sigma^3, \qquad (13)$$

называемую *асимметрией*.

Слайд 24. Пределы значений асимметрии α – от -∞ до + ∞. При $\alpha = 0$ распределение симметрично: $M_0 = \overline{X}$. При $\alpha > 0$ $M_0 < \overline{X}$, а при $\alpha < 0$ $M_0 > \overline{X}$.

Эксцессом называют величину

$$\varepsilon = \mu_4 / \sigma^4 - 3. \qquad (14)$$

Эксцесс показывает степень крутости кривой распределения признака X по сравнению с крутостью нормального распределения, дисперсия которого равна D(x). При $\varepsilon = 0$ распределение нормальное. Если $\varepsilon > 0$, то крутость положительная и кривая распределения имеет более острую вершину, чем при нормальном распределении. Если же $\varepsilon < 0$, то крутость отрицательная и кривая имеет более плоскую вершину. В этом случае возможно даже, что в центре распределения будут выемки (двухмодальная кривая). Значения эксцесса лежат на полусегменте [- 3; + ∞).

Ошибки асимметрии и эксцесса вычисляются соответственно по формулам:

$$E_\alpha = \sqrt{\frac{6 \cdot (n-1)}{(n+1) \cdot (n+2)}}, \qquad (15)$$

$$E_\varepsilon = \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n+1)^2 \cdot (n+3) \cdot (n+5)}}. \qquad (16)$$

Здесь n – объем вариационного ряда.

Пример. Дана статистическая совокупность, характеризующая затраты (в копейках) на рубль продукции (работ, услуг) за 1990 г., по 100 предприятиям г. Минска:

Слайд 25.

61,55	61,59	62,09	63,08	63,97	64,74	65,07
67,12	68,10	69,38	70,21	70,21	70,36	71,25
71,86	72,00	72,39	72,41	72,46	72,50	72,80
72,84	73,44	74,93	75,46	75,65	77,13	77,37
77,64	77,86	77,93	78,03	78,28	78,74	78,97
79,07	79,10	79,34	79,34	79,34	79,40	79,49
79,70	80,02	80,26	80,56	80,65	80,69	81,13
81,32	81,40	81,54	81,85	82,27	82,71	82,74
82,78	83,03	83,05	83,59	83,68	83,74	83,78
83,96	84,98	85,18	85,32	85,64	85,71	85,84
86,01	86,03	86,05	86,11	86,48	86,94	86,98
87,38	87,47	87,59	87,89	88,03	88,04	88,11
88,24	88,98	90,34	90,40	90,58	90,73	90,76
92,51	92,72	92,94	94,58	95,06	95,73	96,11
			96,34	96,55		

Необходимо составить интервальный ряд распределения, вычислить числовые характеристики признака X, характеризующего затраты.

Р е ш е н и е. Каждое индивидуальное измерение затрат представлено отдельно, поэтому их называют *несгруппированными дискретными данными*. Они могут быть подвергнуты группировке в виде дискретного или интервального ряда. Для построения интервального ряда вычислим длину интервала:

$$h = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg n} = \frac{96{,}55 - 61{,}55}{1 + 3{,}322 \cdot \lg 100} \approx 5.$$

В результате ряд примет вид, приведенный в таблице

Слайд 26.

Затраты x_i на 1 руб. продукции, коп.	Число предприятий m_i	Частость W_i	Накопленная частость
61,55 - 66,55	7	0,07	0,07
66,55 – 71,55	7	0,07	0,14
71,55 – 76,55	11	0,11	0,25
76,55 – 81,55	27	0,27	0,52
81,55 – 86,55	23	0,23	0,75
86,55 – 91,55	16	0,16	0,91
91,55 – 96,55	9	0,09	1,00

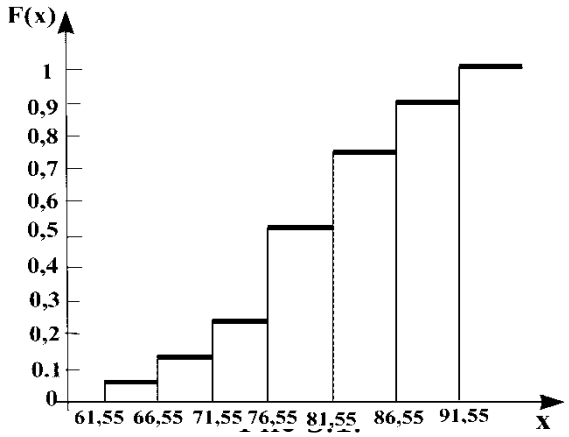
В этой таблице даны частости, вычисленные по формуле $W_i = m_i / n$. Используя накопленные частости $F_i = \sum_{j=1}^{i} W_j$, получаем функцию распределения

$$F(x) = \begin{cases} 0 & при \quad x \le 61{,}55, \\ 0{,}07 & при \quad 61{,}55 < x \le 66{,}55, \\ 0{,}14 & при \quad 66{,}55 < x \le 71{,}55, \\ 0{,}25 & при \quad 71{,}55 < x \le 76{,}55, \\ 0{,}52 & при \quad 76{,}55 < x \le 81{,}55, \\ 0{,}75 & при \quad 81{,}55 < x \le 86{,}55 \\ 0{,}91 & при \quad 86{,}55 < x \le 91{,}55, \\ 1{,}00 & при \quad 91{,}55 < x. \end{cases}$$

Слайд 27. График функции F(x) приведен на рис.1.

Для несгруппированных данных находим среднее арифметическое:

$$\overline{X} = \frac{1}{100} \cdot \sum_{i=1}^{100} x_i = \frac{1}{100} \cdot (61{,}55 + 62{,}09 + \ldots + 96{,}34 + 96{,}55) = 80{,}828 .$$



Если данные представлены в виде интервального ряда, то среднее арифметическое также можно находить по формуле (1): **Слайд 28.**

$$\overline{X} = \frac{1}{100} \cdot \sum_{i=1}^{7} x_i^* m_i = \frac{1}{100} \cdot (64{,}05 \cdot 7 + 69{,}05 \cdot 7 + 74{,}05 \cdot 11 + $$
$$+ 79{,}05 \cdot 27 + 84{,}05 \cdot 23 + 89{,}05 \cdot 16 + 94{,}05 \cdot 9) = 80{,}85 .$$

Здесь x_i^* – среднее значение признака X из интервала $x_i - x_{i+1}$.

Сравнивая полученные средние арифметические, видим, что группировка исходных данных сопровождается потерей точности. Поэтому остальные

числовые характеристики вычислим по несгруппированным данным, которые образуют дискретный вариационный ряд.

С теоретической точки зрения наиболее подходящей мерой колеблемости ряда распределения служит статистическая дисперсия

$$\sigma^2 = \frac{1}{100} \cdot \sum_{i=1}^{100} (x_i - \bar{X})^2 = \frac{1}{100} \cdot (371{,}641 + 370{,}101 + \ldots + 247{,}181) = 70{,}165 .$$

Отсюда $\sigma = \sqrt{70{,}165} = 8{,}376$.

Пределы изменения затрат характеризует размах

$$R = 96{,}55 - 61{,}55 = 35{,}0 .$$

По формуле (12) вычисляем коэффициент вариации: **Слайд 29.**

$$v = \frac{\sigma}{\bar{X}} \cdot 100\% = \frac{8{,}376}{80{,}828} \cdot 100\% = 10{,}36 .$$

Величина этого коэффициента показывает, что совокупность исходных данных однородна.

Выяснение общего характера распределения предполагает вычисление асимметрии и эксцесса соответственно по формулам (5.13) и (5.14):

$$\alpha = \frac{1}{100 \cdot \sigma^3} \sum_{i=1}^{100} (x_i - \bar{X})^3 = \frac{-3373{,}1453}{100 \cdot 578{,}7334} = -0.0574 ,$$

$$\varepsilon = \frac{1}{100 \cdot \sigma^4} \sum_{i=1}^{100} (x_i - \bar{X})^4 - 3 = \frac{1317212{,}245}{492311{,}4} - 3 = -0{,}3244 .$$

Асимметрия отрицательна, следовательно, распределение характеризуется незначительной левосторонней асимметрией. Отрицательный эксцесс указывает на более плосковершинное распределение по сравнению с нормальным.

Ошибки асимметрии и эксцесса находим по формулам (15), (16):

$$E_\alpha = \sqrt{\frac{6 \cdot (100 - 1)}{(100 + 1) \cdot (100 + 3)}} = 0{,}2389 , \textbf{ Слайд 30.}$$

$$E_\varepsilon = \sqrt{\frac{24 \cdot 100 \cdot (100 - 2) \cdot (100 - 3)}{(100 + 1)^2 \cdot (100 + 3) \cdot (100 + 5)}} = 0{,}4547 .$$

Так как отношения $|\alpha| / E_\alpha = 0{,}0574/02389 = 0{,}24$ и $|\varepsilon| / E_\varepsilon = 0{,}3244/0{,}4547 = 0{,}71$ меньше числа 3, то асимметрия и эксцесс незначительны в распределении затрат.

5.2.6. Точечные оценки параметров распределения случайной величины

Рассмотрим задачу об определении неизвестных параметров, от которых зависит закон распределения случайной величины, по ограниченному числу опытов, т. е. выборке. Любое значение искомого параметра, вычисленное на основе ограниченного числа опытов, всегда будет содержать элемент случайности. Такое приближенное, случайное значение будем называть *оценкой*.

Статистические оценки делятся на точечные и интервальные. Оценка, определяемая одним числом, называется *точечной*.

Пусть X – случайная величина, закон распределения которой содержит неизвестный параметр a. Обозначим наблюдаемые значения случайной величины $x_1, x_2, ..., x_n$. Их можно рассматривать как n независимых случайных величин, каждая из которых распределена по тому же закону, что и случайная величина X. Требуется найти подходящую оценку для параметра a по результатам n независимых опытов.

Слайд 31. Обозначим через $\tilde{a}$ оценку для параметра a. Она должна представлять собой функцию величин $x_1, x_2, ..., x_n$:

$$\tilde{a} = \tilde{a}\,(x_1, x_2, ..., x_n).$$

К оценке $\tilde{a}$ предъявляется ряд требований:

1. Чтобы, пользуясь величиной $\tilde{a}$ вместо a, не делать систематических ошибок в сторону занижения или завышения, т. е. чтобы $M(\tilde{a}) = a$. Оценка, удовлетворяющая такому условию, называется *несмещенной*. Величина смещения определяется по формуле $b(\tilde{a}) = M(\tilde{a}) - a$.

2. Чтобы с увеличением числа опытов n случайная величина $\tilde{a}$ приближалась (сходилась по вероятности) к параметру a, т. е. чтобы $D(\tilde{a}) \to 0$ или

$$\lim_{n\to\infty} (P(\tilde{a} - a) \geq \varepsilon) = 0.$$

Оценка, обладающая таким свойством, называется *состоятельной*.

3. Чтобы выбранная несмещенная оценка обладала по сравнению с другими наименьшей дисперсией, т. е.

$$D(\tilde{a}) = \min.$$

Оценка, обладающая таким свойством, называется *эффективной*.

Слайд 32. В качестве *оценки для математического ожидания* предлагается брать среднее арифметическое наблюдаемых значений:

$$\tilde{m} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \, .$$

Эта оценка является состоятельной и несмещенной.

Несмещенной оценкой дисперсии является величина

$$\tilde{D} = \frac{1}{n-1} \cdot \sum_{i-1}^{n} (x_i - \tilde{m})^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \tilde{m}^2 \right) \cdot \frac{n}{n-1} \, .$$

Замечание. Сравнивая формулы

$$\sigma^2 = \frac{\sum_{i=1}^{m} n_i (x_i - \bar{x})^2}{n} \text{ и } \tilde{D} = \frac{\sum_{i=1}^{m} n_i (x_i - \bar{x})^2}{n-1}$$

видим, что они отличаются лишь знаменателями. Очевидно, при достаточно больших значениях n объема выборки, выборочная и исправленная дисперсия (несмещенная оценка) будут мало различаться. На практике пользуются исправленной дисперсией, если $n < 30$.

Эффективность или неэффективность оценки зависит от вида закона распределения величины X .

Одним из важнейших методов нахождения оценок параметров распределения по данным выборки является метод максимального правдоподобия.

Слайд 33. Метод максимального правдоподобия

Метод максимального правдоподобия является широко распространенным методом точечной оценки. Он предложен в 1912 г. английским статистиком Р. Фишером.

Пусть из генеральной совокупности с плотностью распределения вероятностей $f(x,a)$ произведена выборка объема n и получены результаты $x_1, x_2, ..., x_n$. Предположим вначале, что X - дискретная случайная величина, закон распределения которой зависит от неизвестного параметра a. Например, можно предположить, что случайная величина X имеет распределение Пуассона $P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$, где $a = \lambda$ - неизвестный параметр, который надо оценить по данным выборки. Будем рассматривать результаты выборки как реализацию n – мерной случайной величины $(X_1, X_2, ..., X_n)$. Предположим далее, что составляющие этой случайной величины независимы. В этом случае вероятность того, что составляющие примут значения, равные наблюдаемым, (она называется функцией правдоподобия) равна

$$L = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(x_1, a) \cdot P(x_2, a) \cdot ... \cdot P(x_n, a) = \prod_{i=1}^{n} P(x_i, a) \, .$$

В случае непрерывной случайной величины функция правдоподобия имеет вид

$$L = f(x_1, x_2, ..., x_n) = f(x_1, a) \cdot f(x_2, a) \cdot ... \cdot f(x_n, a) = \prod_{i=1}^{n} f(x_i, a). \qquad (19)$$

Слайд 34. Формула (19) определяет плотность распределения вероятностей непрерывной случайной величины $(X_1, X_2, ..., X_n)$, или плотность распределения выборки.

В качестве оценки неизвестного параметра a, найденной по методу максимального правдоподобия, выбирается такая функция $\tilde{a} = u(x_1, x_2, ..., x_n)$, которая максимизирует функцию правдоподобия. Следовательно, на основании известных правил дифференциального исчисления для нахождения оценок максимального правдоподобия составляется система m уравнений (m – число оцениваемых параметров):

$$\frac{\partial L}{\partial a_i} = 0 \quad (i = 1,2,3,..., m)$$

и выбирается то решение, которое обращает функцию правдоподобия в максимум. Поскольку экстремум функции L и $\ln L$ достигается при одних и тех же значениях $\tilde{a} = u(x_1, x_2, ..., x_n)$, то иногда для упрощения расчетов пользуются логарифмической функцией правдоподобия. В этом случае оценки максимального правдоподобия находятся из системы уравнений $\frac{\partial \ln L}{\partial a_i} = 0 \quad (i = 1,2,..., m)$.

Слайд 35. Метод максимального правдоподобия обладает рядом преимуществ по сравнению с методом моментов.

Укажем некоторые важные свойства оценок максимального правдоподобия:

1) метод максимального правдоподобия дает состоятельные оценки;
2) если существует эффективная оценка, то метод максимального правдоподобия дает эту оценку;
3) оценки максимального правдоподобия асимптотически эффективны;
4) оценки максимального правдоподобия имеют асимптотически нормальное распределение с параметрами

$$M(\tilde{a}) = a, \quad D(\tilde{a}) = -\frac{1}{M\left(\frac{\partial^2 \ln f(x,a)}{\partial a^2}\right)};$$

3) если существуют достаточные оценки, то метод максимального правдоподобия дает их.

Недостаток метода заключается в том, что иногда оценки максимального правдоподобия являются смещенными. Смещение можно устранить введением поправок (с ростом n смещение уменьшается, т. е. оценки максимального правдоподобия асимптотически несмещенные). Кроме того, для нахождения оценок методом максимального правдоподобия приходится решать сложные системы уравнений.

Для самостоятельного изучения

Слайд 36. Интервальные оценки параметров распределения случайной величины

В ряде задач требуется не только найти для параметра a подходящее численное значение, но и оценить его точность и надежность. Требуется знать – к каким ошибкам может привести замена параметра a его точечной оценкой $\tilde{a}$ и с какой степенью уверенности можно ожидать, что эти ошибки не выйдут за известные пределы?

Такого рода задачи особенно актуальны при малом числе наблюдений, когда точечная оценка $\tilde{a}$ в значительной мере случайна и приближенная замена a на $\tilde{a}$ может привести к серьезным ошибкам.

Чтобы дать представление о точности и надежности оценки $\tilde{a}$, в математической статистике пользуются так называемыми *доверительным и интервалами* и *доверительными вероятностями*.

Пусть для параметра a получена на опыте несмещенная оценка $\tilde{a}$. Мы хотим оценить возможную при этом ошибку. Назначим некоторую достаточно большую вероятность β (например, $\beta = 0{,}9,$ или $0{,}99$) такую, что событие с вероятностью β можно считать практически достоверным, и найдем такое значение ε, для которого

$$P(|a - \tilde{a}| < \varepsilon) = \beta. \qquad (20)$$

Слайд 37. Тогда диапазон практически возможных значений ошибки, возникающий при замене a на $\tilde{a}$, будет $\pm\varepsilon$; большие по абсолютной величине ошибки будут появляться только с малой вероятностью $\alpha = 1 - \beta$.

Ясно, что $\tilde{a}$ тем точнее определяет параметр a, чем меньше абсолютная величина разности $|a - \tilde{a}|$, т. е., если $\varepsilon > 0$ и $|a - \tilde{a}| < \varepsilon$, то, чем меньше ε, тем оценка точнее. Таким образом, положительное число ε характеризует *точность оценки*. Кроме того, статистические методы не позволяют категорически утверждать, что оценка $\tilde{a}$ удовлетворяет неравенству

$|a-\tilde{a}|<\varepsilon$; можно лишь говорить о вероятности β, с которой это неравенство осуществляется.

Перепишем (20) в виде:

$$P(\tilde{a}-\varepsilon<a<\tilde{a}+\varepsilon)=\beta. \qquad (21)$$

Равенство (21) означает, что с вероятностью β неизвестное значение параметра a попадает в интервал

$$I_\beta=(\tilde{a}-\varepsilon;\tilde{a}+\varepsilon). \qquad (22)$$

При этом необходимо отметить одно обстоятельство. Ранее мы неоднократно рассматривали вероятность попадания случайной величины в заданный неслучайный интервал. Здесь дело обстоит иначе: величина a *не случайна*, зато случаен интервал I_β. Случайно его положение на оси абсцисс, определяемое его центром $\tilde{a}$; случайна вообще и длина интервала 2ε, так как величина ε вычисляется, как правило, по опытным данным. Поэтому в данном случае лучше будет толковать величину β не как вероятность "попадания" точки a в интервал I_β, а как вероятность того, что случайный интервал I_β *накроет* точку a (рис. 2).

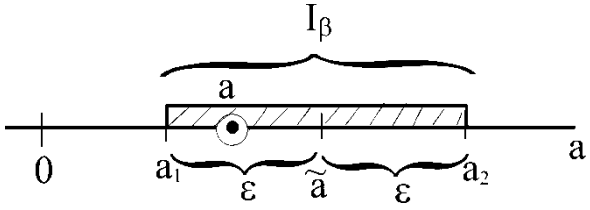


Рис. 2

Слайд 38. Вероятность β принято называть *доверительной вероятностью (надежностью)*, а интервал I_β – *доверительным интервалом*. [1]). Границы интервала I_β: $a_1=\tilde{a}-\varepsilon$ и $a_2=\tilde{a}+\varepsilon$ называются доверительными границами.

Если бы нам был известен закон распределения величины $\tilde{a}$, задача нахождения доверительного интервала была бы весьма проста: достаточно было бы найти такое значение ε, для которого

$$P(|\tilde{a}-a|<\varepsilon)=\beta.$$

Затруднение состоит в том, что закон распределения оценки $\tilde{a}$ зависит от закона распределения величины X и, следовательно, от его неизвестных параметров (в частности, и от самого параметра a).

---

[1]) На рис. 5.2 рассматривается доверительный интервал, симметричный относительно $\tilde{a}$. Вообще говоря, это не обязательно.

1. Рассмотрим задачу о доверительном интервале для оценки математического ожидания нормального распределения при известной дисперсии.

Пусть проведено n независимых опытов над случайной величиной X, характеристики которой – математическое ожидание m и дисперсия D неизвестны. Для этих параметров получены оценки: **Слайд 39.**

$$\tilde{m} = \frac{\sum_{i=1}^{n} X_i}{n}; \quad \tilde{D} = \frac{\sum_{i=1}^{n} (X_i - \tilde{m})^2}{n-1}. \tag{23}$$

Требуется построить доверительный интервал I_β, соответствующий доверительной вероятности β, для математического ожидания m величины X.

При решении этой задачи воспользуемся тем, что величина $\tilde{m}$ представляет собой сумму n независимых одинаково распределенных случайных величин X_i, и, согласно центральной предельной теореме, при достаточно большом n ее закон распределения близок к нормальному. На практике же даже при относительно небольшом числе слагаемых (порядка 10–20) закон распределения суммы можно приближенно считать нормальным. Будем исходить из того, что величина $\tilde{m}$ распределена по нормальному закону. Характеристики этого закона – математическое ожидание и дисперсия – равны соответственно m и $\frac{D}{n}$. Предположим, что величина D нам известна, и найдем такую величину ε_β, для которой

$$P(|\tilde{m} - m| < \varepsilon_\beta) = \beta. \tag{24}$$

Слайд 40. Выразим вероятность в левой части (24) через нормальную функцию распределения

$$P(|\tilde{m} - m| < \varepsilon_\beta) = 2 \cdot \Phi\left(\frac{\varepsilon_\beta}{\sigma_{\tilde{m}}}\right), \tag{25}$$

где $\sigma_{\tilde{m}} = \sqrt{\frac{D}{n}}$ – среднее квадратическое отклонение оценки $\tilde{m}$.

Из уравнения $2 \cdot \Phi\left(\frac{\varepsilon_\beta}{\sigma_{\tilde{m}}}\right) = \beta$ находим значение ε_β:

$$\varepsilon_\beta = \sigma_{\tilde{m}} \cdot \arg\Phi\left(\frac{\beta}{2}\right), \tag{26}$$

где $\Phi(x)$ – функция Лапласа (приложение 2), а $\arg\Phi(x)$ – обратная ей функция, т. е. такое значение аргумента, при котором нормальная функция распределения равна x.

Слайд 41. Дисперсия D, через которую выражена величина $\sigma_{\tilde{m}}$, нам в точности не известна; в качестве ее ориентировочного значения можно воспользоваться оценкой $\tilde{D}$ (23) и положить приближенно:

$$\sigma_{\tilde{m}} = \sqrt{\frac{\tilde{D}}{n}}.$$

Таким образом, приближенно решена задача построения доверительного интервала, который равен:

$$I_\beta = (\tilde{m} - \varepsilon_\beta;\ \tilde{m} + \varepsilon_\beta),$$

где ε_β определяется формулой (26).

Смысл полученного соотношения таков: с надежностью β можно утверждать, что доверительный интервал

$$\left(\tilde{m} - \sigma_{\tilde{m}} \cdot \arg\Phi\left(\frac{\beta}{2}\right);\ \tilde{m} + \sigma_{\tilde{m}} \cdot \arg\Phi\left(\frac{\beta}{2}\right)\right)$$

покрывает неизвестный параметр m; точность оценки $\varepsilon_\beta = \sigma_{\tilde{m}} \cdot \arg\Phi\left(\frac{\beta}{2}\right)$.

Пример. Случайная величина X имеет нормальное распределение с известным средним квадратическим отклонением $\sigma = 3$. Найти доверительные интервалы для оценки неизвестного математического ожидания m по выборочным средним $\bar{x}$, если объем выборки $n = 36$ и задана надежность оценки $\beta = 0{,}95$.

Слайд 42. Р е ш е н и е. Из соотношения

$$2\Phi\left(\frac{\varepsilon_\beta}{\sigma_{\tilde{m}}}\right) = 0{,}95 \text{ получим } \Phi\left(\frac{\varepsilon_\beta}{\sigma_{\tilde{m}}}\right) = 0{,}475\cdot$$

По таблице (приложение 2) найдем

$$\frac{\varepsilon_\beta}{\sigma_{\tilde{m}}} = 1{,}96, \text{ где } \sigma_{\tilde{m}} = \frac{\sigma}{\sqrt{n}},\ \tilde{m} = \bar{x}.$$

Найдем точность оценки:

$$\varepsilon_\beta = \frac{\sigma \cdot \arg\Phi\left(\frac{\beta}{2}\right)}{\sqrt{n}} = \frac{1{,}96 \cdot 3}{\sqrt{36}} = 0{,}98.$$

Запишем доверительный интервал:

$$(\tilde{m} - 0{,}98;\ \tilde{m} + 0{,}98).$$

Например, если $\bar{x} = \tilde{m} = 4{,}1$, то доверительный интервал имеет следующие доверительный границы:

$$\tilde{m} - 0{,}98 = 4{,}1 - 0{,}98 = 3{,}12;$$
$$\tilde{m} + 0{,}98 = 4{,}1 + 0{,}98 = 5{,}08.$$

Слайд 43. Таким образом, значения неизвестного параметра m, согласующиеся с данными выборки, удовлетворяют неравенству

$$3{,}12 < m < 5{,}08.$$

Поясним смысл, который имеет заданная надежность $\beta = 0{,}95$. Она указывает, что если произведено достаточно большое число выборок, то 95% из них определяют такие доверительные интервалы, в которых параметр действительно заключен; лишь в 5% случаев он может выйти за границы доверительного интервала.

Замечание. Минимальный объем выборки, который обеспечит требуемую оценку математического ожидания с достаточной точностью ε и надежностью β, можно найти по формуле

$$n = \frac{\left[\arg \Phi\left(\frac{\beta}{2}\right)\right]^2 \cdot \sigma^2}{\varepsilon^2}$$

2. Рассмотрим задачу о доверительном интервале для математического ожидания нормального распределения при неизвестной дисперсии.

Слайд 44. Необходимо найти доверительный интервал, покрывающий математическое ожидание m нормально распределенной случайной величины X с заданной надежностью β.

По выборке $x_1, x_2, ..., x_n$ найдем среднее арифметическое $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ и дисперсию $\tilde{D} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Рассмотрим новую случайную величину

$$t = \frac{\bar{x} - m}{\sigma_{\tilde{m}}},$$

где $\sigma_{\tilde{m}} = \sqrt{\frac{\tilde{D}}{n}}$, $\bar{x} = \tilde{m}$. Функция распределения этой случайной величины называется законом распределения Стьюдента (или $t-$ распределением) с $(n-1)$ степенями свободы. В таблице (приложение 4) с двумя входами дается

решение уравнения $P(|t|>x)=q$. Строка выбирается по числу $(n-1)$ степеней свободы, а столбец – по значению q. На пересечении выбранных строки и столбца находится число x.

Слайд 45. Чтобы получить границы доверительного интервала для параметра m с надежностью β, необходимо найти по таблице такое число t_β, для которого выполняется неравенство

$$P(|t|<t_\beta)=\beta.$$

Воспользуемся тем, что события $|t|\le t_\beta$ и $|t|>t_\beta$ противоположны. Если $P(|t|>t_\beta)=q$, а $P(|t|<t_\beta)=\beta$, то

$$\beta=P(|t|<t_\beta)=1-P(|t|>t_\beta)=1-q.$$

Следовательно, $q=1-\beta$. По числу степеней свободы $(n-1)$ и числу $q=1-\beta$ находим по таблице t_β. Неравенство

$$\left|\frac{\bar{x}-m}{\sigma_{\tilde{m}}}\right|<t_\beta$$

преобразуется в равносильное ему неравенство

$$-t_\beta<\frac{m-\bar{x}}{\sigma_{\tilde{m}}}<t_\beta;$$

$$-t_\beta\cdot\sigma_{\tilde{m}}<m-\bar{x}<t_\beta\cdot\sigma_{\tilde{m}};$$

$$\bar{x}-t_\beta\cdot\sigma_{\tilde{m}}<m<\bar{x}+t_\beta\cdot\sigma_{\tilde{m}}.$$

Или, учитывая, что $\bar{x}=\tilde{m}$, запишем искомый доверительный интервал:

Слайд 46.

$$(\tilde{m}-t_\beta\cdot\sigma_{\tilde{m}}<m<\tilde{m}+t_\beta\cdot\sigma_{\tilde{m}}).$$

Пример. Количественный признак X генеральной совокупности распределен нормально. По выборке объема $n=16$ найдены выборочная средняя $\bar{x}=20{,}2$ и исправленное среднее квадратическое отклонение $\sqrt{\tilde{D}}=0{,}8$. Оценить неизвестное математическое ожидание при помощи доверительного интервала с надежностью $\beta=0{,}95$.

Р е ш е н и е. Пользуясь таблицей (приложение 4), по $\beta=0{,}95$ и $n=16$ находим $t_\beta=2{,}13$. Найдем доверительные границы:

$$\bar{x} - t_\beta \cdot \sigma_{\tilde{m}} = 20{,}2 - 2{,}13 \cdot \frac{0{,}8}{\sqrt{16}} = 19{,}774,$$

$$\bar{x} + t_\beta \cdot \sigma_{\tilde{m}} = 20{,}2 + 2{,}13 \cdot \frac{0{,}8}{\sqrt{16}} = 20{,}626.$$

Итак, с надежностью 0,95 неизвестный параметр m заключен в доверительный интервал $19{,}774 < m < 20{,}626$.

Замечание 1. При неограниченном возрастании объема выборки n распределение Стьюдента стремится к нормальному. Поэтому при $n > 30$ можно вместо распределения Стьюдента пользоваться нормальным распределением. **Слайд 47.** Однако для малых выборок $(n < 30)$ замена распределения нормальным приводит к грубым ошибкам, а именно – к неоправданному сужению доверительного интервала, т. е. к повышению точности оценки. Например, если $n = 5$ и $\beta = 0{,}99$, то, пользуясь распределением Стьюдента, найдем $t_\beta = 4{,}6$, используя функцию Лапласа, найдем $t_\beta = 2{,}58$, т. е. доверительный интервал в последнем случае окажется более узким, чем найденный по распределению Стьюдента. Это вовсе не свидетельствует о слабости метода Стьюдента, а объясняется тем, что малая выборка содержит малую информацию об интересующем нас признаке.

Замечание 2. Если Z – нормальная величина с $m_z = 0$ и $\sigma_z = 1$, а V – независимая от z величина, распределенная по закону χ^2 с k степенями свободы, то величина

$$T = \frac{Z}{\sqrt{\frac{V}{k}}} \qquad (27)$$

распределена по закону Стьюдента с k степенями свободы.

Слайд 48. Пусть количественный признак X генеральной совокупности распределен нормально с $m_x = a, \sigma_x = \sigma$. Если из этой совокупности извлекать выборки объема n и по ним находить выборочные средние $\bar{x}$, то можно доказать, что выборочная средняя распределена нормально, причем

$$m_x = a, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Тогда величина

$$Z = \frac{\bar{x} - a}{\frac{\sigma}{\sqrt{n}}} \qquad (28)$$

также имеет нормальное распределение, как линейная функция нормального аргумента $\bar{x}$, причем, $m_z = 0, \sigma_z = 1$. Доказано, что независимая от Z случайная величина

$$V = \frac{(n-1)\cdot \tilde{D}}{\sigma^2} \qquad (29)$$

($\tilde{D}$ – исправленная выборочная дисперсия) распределена по закону χ^2 с $k = n-1$ степенями свободы.

Слайд 49. Следовательно, подставив (28) и (29) в (27), получим величину

$$T = \frac{(\bar{x} - a)\sqrt{n}}{\sqrt{\tilde{D}}},$$

которая распределена по закону Стьюдента с $k = n-1$ степенями свободы.

3. Рассмотрим задачу о доверительном интервале для оценки среднего квадратического отклонения σ нормального распределения.

Пусть количественный признак X генеральной совокупности распределен нормально. Требуется оценить неизвестное генеральное среднее квадратическое отклонение σ по “исправленному” выборочному среднему квадратическому отклонению $\sigma = \sqrt{\tilde{D}}$. Найдем доверительные интервалы, покрывающие параметр σ с заданной надежностью β.

Потребуем, чтобы выполнялось соотношение

$$P\left(\left|\sigma - \sqrt{\tilde{D}}\right| < \varepsilon\right) = \beta$$

или

$$P\left(\sqrt{\tilde{D}} - \varepsilon < \sigma < \sqrt{\tilde{D}} + \varepsilon\right) = \beta.$$

Для того, чтобы можно было пользоваться готовой таблицей, преобразуем двойное неравенство

$$\sqrt{\tilde{D}} - \varepsilon < \sigma < \sqrt{\tilde{D}} + \varepsilon$$

в равносильное неравенство **Слайд 50.**

$$\sqrt{\tilde{D}}\left(1 - \frac{\varepsilon}{\sqrt{\tilde{D}}}\right) < \sigma < \sqrt{\tilde{D}}\left(1 + \frac{\varepsilon}{\sqrt{\tilde{D}}}\right).$$

Положив $\frac{\varepsilon}{\sqrt{\tilde{D}}} = q$, получим

$$\sqrt{\tilde{D}}\,(1-q) < \sigma < \sqrt{\tilde{D}}\,(1+q). \qquad (30)$$

Для того, чтобы найти q, введем в рассмотрение случайную величину χ:

$$\chi = \frac{\sqrt{\tilde{D}}}{\sigma}\sqrt{n-1},$$

где n – объем выборки. Как было указано (замечание 2, соотношение (29)), величина $\frac{\tilde{D}(n-1)}{\sigma^2}$ распределена по закону χ^2, поэтому корень из нее обозначают χ.

Преобразуем неравенство (30) так, чтобы оно приняло вид

$$\chi_{\min} < \chi < \chi_{\max}.$$

Вероятность этого неравенства равна заданной вероятности β. Предполагая, что $q < 1$, перепишем (30): **Слайд 51.**

$$\frac{1}{\sqrt{\tilde{D}}(1+q)} < \frac{1}{\sigma} < \frac{1}{\sqrt{\tilde{D}}(1-q)}.$$

Умножив все члены неравенства на $\sqrt{\tilde{D}(n-1)}$, получим

$$\frac{\sqrt{n-1}}{1+q} < \frac{\sqrt{\tilde{D}(n-1)}}{\sigma} < \frac{\sqrt{n-1}}{1-q} \quad \text{или} \quad \frac{\sqrt{n-1}}{1+q} < \chi < \frac{\sqrt{n-1}}{1-q}.$$

Вероятность того, что это неравенство, а следовательно, и равносильное ему неравенство (30) будет выполняться, равна β.

Практически для отыскания q пользуются таблицей (приложение 4). Вычислив по выборке $\sqrt{\tilde{D}}$ и найдя по таблице q, получим искомый доверительный интервал (30), покрывающий σ с заданной надежностью β, т.е. интервал $\sqrt{\tilde{D}}(1-q) < \sigma < \sqrt{\tilde{D}}(1+q)$.

Пример. Количественный признак X генеральной совокупности распределен нормально. По выборке объема $n = 25$ найдена несмещенная оценка $\tilde{D} = 0{,}64$. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение σ с надежностью 0,95.

Слайд 52. Р е ш е н и е. По таблице (приложение 4) для $\beta = 0{,}95$ и $n = 25$ найдем $q = 0{,}32$. Искомый доверительный интервал (30) таков:

$$0{,}8(1-0{,}32) < \sigma < 0{,}8(1+0{,}32) \quad \text{или} \quad 0{,}544 < \sigma < 1{,}056.$$