ЧАСТНОЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ «СЕВЕРО-КАВКАЗСКИЙ КОЛЛЕДЖ ИННОВАЦИОННЫХ ТЕХНОЛОГИЙ»

ЛЕКЦИЯ 14. ТЕМА 13. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1. Предмет математической статистики

Математическая статистика — раздел математики, изучающий методы сбора, систематизации, обработки и интерпретации статистических данных с целью выявления закономерностей, свойственных массовым случайным явлениям. В отличие от теории вероятностей, которая по заданным вероятностям предсказывает поведение случайных величин, математическая статистика решает обратную задачу: по результатам наблюдений восстанавливает (оценивает) характеристики изучаемого распределения.

Центральное понятие статистики — **статистическая совокупность** — множество однородных объектов, объединённых по какому-либо признаку. Отдельный объект называется единицей совокупности, а интересующая исследователя характеристика — варьирующим (изменяющимся) признаком.

2. Генеральная совокупность и выборка

Генеральная совокупность — полное множество всех объектов, относительно которых делаются статистические выводы. Так, при изучении роста десятиклассников региона генеральной совокупностью выступают все десятиклассники региона.

Поскольку сплошное обследование (перепись) часто невозможно или нецелесообразно — из-за стоимости, времени, либо разрушающего характера испытаний (например, испытание каждой лампы на долговечность), — исследование проводят по **выборке**: отобранной из генеральной совокупности части объектов.

Ключевое требование к выборке — **репрезентативность**: выборка должна правильно отражать структуру и свойства генеральной совокупности. Это достигается рандомизацией (случайным отбором), при котором каждая единица генеральной совокупности имеет определённую вероятность попасть в выборку. Наиболее распространены простая случайная выборка (жеребьёвка, генератор случайных чисел) и стратифицированная выборка (разбиение совокупности на однородные группы с отбором из каждой).

Число единиц генеральной совокупности и выборки обозначают соответственно N и n. Репрезентативная выборка позволяет по её характеристикам с известной надёжностью судить о характеристиках генеральной совокупности — в этом состоит основная идея выборочного метода.

3. Представление статистических данных

3.1. Вариационный ряд и таблицы

Результаты наблюдений образуют **вариационный ряд** — упорядоченный по возрастанию ряд значений признака. Если значение встречается несколько раз, оно называется вариантой, а число её повторений — частотой (относительная частота — отношение частоты к объёму выборки).

Первичная систематизация данных осуществляется с помощью **таблиц распределения частот**. Например, при наблюдении числа успехов в серии из 10 испытаний таблица выглядит так:

Значение x_i	0	1	2	3	4	5	…
Частота n_i	3	8	14	22	30	23	…

Накопленная частота показывает число наблюдений, не превосходящих данного значения; она служит основой для построения эмпирической функции распределения.

3.2. Графические представления

Основные формы наглядного изображения данных:

• Столбчатая диаграмма — изображение частот прямоугольниками, высоты которых пропорциональны значениям признака; применяется для дискретных и категориальных данных.

• Полигон частот — ломаная, соединяющая точки с координатами (xi, ni)(*xi*,*ni*); используется для дискретных рядов.

• Гистограмма — над каждым интервалом значений признака строится прямоугольник, площадь которого пропорциональна частоте попадания в интервал; основной инструмент для непрерывных данных.

• Круговая диаграмма — отражает доли категорий в общей совокупности.

• Полигон накопленных частот (кумулята) и ящик с усами (boxplot) — последние отображают медиану, квартили и выбросы.

• Диаграмма рассеяния — визуализирует связь между двумя количественными признаками.

Выбор вида диаграммы определяется типом данных: категориальные признаки удобно представлять столбчатыми и круговыми диаграммами, количественные — гистограммами и полигонами.

4. Обобщающие (числовые) характеристики

4.1. Меры центральной тенденции

Среднее арифметическое выборки объёма n со значениями $x_1, x_2, \ldots, x_n$:

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

Среднее — единственная мера центра, использующая все значения наблюдений. Её существенный недостаток — высокая чувствительность к выбросам: одно аномальное значение способно существенно сместить среднее.

Медиана — значение, делящее упорядоченный ряд пополам: при нечётном nn — центральный элемент, при чётном — полусумма двух центральных элементов. Медиана устойчива к выбросам, поэтому при асимметричных распределениях (например, доходах населения) она информативнее среднего.

Мода — наиболее часто встречающееся значение; единственная мера центра, применимая к категориальным данным.

Соотношение среднего, медианы и моды служит диагностическим признаком формы распределения: при симметричном распределении они совпадают; если среднее больше медианы, распределение имеет правую асимметрию (вытянут «правый хвост»), если меньше — левую.

4.2. Меры рассеяния

Размах — разность наибольшего и наименьшего значений; грубая, но простая характеристика изменчивости.

Выборочная дисперсия:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Выборочное стандартное отклонение $S = \sqrt{S^2}$ выражает средний масштаб отклонения значений от среднего и измеряется в тех же единицах, что и сам признак. При оценке дисперсии генеральной совокупности по выборке используется исправленная (несмещённая) оценка с делением на $n - 1$.

Для приближённого описания формы распределения применяется **правило трёх сигм**: при близости распределения к нормальному в интервале $\bar{x} \pm S$ лежит около 68% наблюдений, в интервале $\bar{x} \pm 2S$ — около 95%, в интервале $\bar{x} \pm 3S$ — около 99,7%.

Дополняют картину **квартили** (значения, делящие упорядоченный ряд на четыре части) и **межквартильный размах** — разность третьей и первой квартилей, устойчивая к выбросам мера разброса.

5. Задачи математической статистики

Содержание статистической теории задаётся тремя классами задач.

Описательная статистика — систематизация и наглядное представление данных: построение таблиц, гистограмм, вычисление выборочных характеристик.

Статистическое оценивание — восстановление свойств генеральной совокупности по выборке. Различают точечные оценки (выборочное среднее как оценка математического ожидания, исправленная выборочная дисперсия как оценка дисперсии) и интервальные оценки. Базовый инструмент интервального оценивания — **доверительный интервал**: интервал, который с заранее заданной вероятностью (уровнем доверия, обычно 0,95) накрывает истинное значение параметра. Для среднего при известной дисперсии он имеет вид

$$\bar{x} \pm z\,\frac{\sigma}{\sqrt{n}},$$

где z — квантиль нормального распределения (при уровне 0,95 $z \approx 1{,}96$). Из формулы видно фундаментальное соотношение: ошибка оценивания убывает пропорционально $\sqrt{n}$, то есть повышение точности в 10 раз требует увеличения объёма выборки в 100 раз.

Проверка статистических гипотез — формализованная процедура принятия решений о свойствах генеральной совокупности. Выдвигается нулевая гипотеза H_0 (например, «средние значения в двух группах равны») и по выборочным данным вычисляется статистика, значение которой сравнивается с критической границей. Результат выражается через уровень значимости α — вероятность ошибочно отвергнуть верную нулевую гипотезу (ошибка I рода); стандартные значения $\alpha = 0{,}05$ или $\alpha = 0{,}01$. Важно понимать: утверждение «гипотеза отвергнута на уровне 0,05» означает, что наблюдаемые данные маловероятны при условии справедливости H_0, а не то, что H_0 доказанно ложна.

6. Связь с теорией вероятностей. Решение практических задач

Теоретической основой статистики служат понятия теории вероятностей: случайная величина, закон распределения, математическое ожидание $E(X)$ и дисперсия $D(X)$. Выборочные характеристики при этом играют роль приближённых значений соответствующих теоретических величин; точность этого приближения регулируется законом больших чисел и центральной предельной теоремой, согласно которой сумма большого числа независимых случайных слагаемых имеет распределение, близкое к нормальному, — именно это обосновывает применение нормального распределения при оценивании средних.

Рассмотрим типовые модели практических задач.

Задача 1 (оценка среднего). При контроле качества отобрано $n = 25$ деталей; среднее отклонение размера $\bar{x} = 2{,}0$ мм, выборочное стандартное отклонение $s = 0{,}5$ мм. При уровне доверия 0,95 доверительный интервал для среднего отклонения по всей партии имеет вид

$$2{,}0 \pm 1{,}96 \cdot \frac{0{,}5}{\sqrt{25}} = 2{,}0 \pm 0{,}2,$$

то есть от 1,8 до 2,2 мм. Если технологический допуск равен 2,0 ± 0,3 мм, интервал целиком лежит внутри допуска, и партия признаётся соответствующей требованиям.

Задача 2 (проверка гипотезы). Утверждается, что новый метод обучения повышает средний балл теста. В контрольной группе средний балл 65, в экспериментальной — 69 при сопоставимых объёмах групп. Для проверки рассчитывается статистика разности средних, её распределение при справедливости H_0 («метод не действует») сравнивается с критическим значением; если расчётное значение превышает критическое, различие признаётся статистически значимым на выбранном уровне. При этом следует учитывать, что статистическая значимость не тождественна практической: при очень больших выборках статистически значимыми окажутся и практически пренебрежимо малые различия.

Задача 3 (вероятностный расчёт). Вероятность того, что случайно выбранное изделие дефектно, оценена по выборке как 0,02. Вероятность того, что среди 10 независимых изделий дефектных окажется ровно одно, вычисляется по формуле Бернулли:

$$P = C_{10}^{1}\,(0{,}02)^{1}\,(0{,}98)^{9} \approx 0{,}167.$$

Аналогично по биномиальному распределению решаются задачи контроля качества, планирования ресурсов, страховых расчётов.

Задача 4 (регрессия и прогноз). По диаграмме рассеяния и коэффициенту корреляции (мера линейной связи, значения от −1 до 1) строятся уравнения регрессии, позволяющие прогнозировать значение одного признака по значениям других. Применение регрессии корректно лишь в пределах наблюдавшегося диапазона данных: экстраполяция за его границы — типичный источник ошибочных выводов.

7. Методологические замечания

При применении статистических методов следует учитывать типичные ошибки. Во-первых, **корреляция не означает причинности**: связь между признаками может объясняться общей причиной или случайным совпадением. Во-вторых, выводы по нерепрезентативной выборке (например, по опросу добровольцев) не переносятся на генеральную совокупность. В-третьих, множественные сравнения при большом числе проверяемых гипотез приводят к ложным «открытиям»: при уровне 0,05 в среднем одна из двадцати проверок случайно даст значимый результат. Наконец, любые статистические выводы опираются на предположения о характере данных; их проверка — обязательная часть анализа.