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

ЛЕКЦИЯ 5. ТЕМА 5. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

1. Задачи и методы математической статистики. Виды выборки

Математическая статистика представляет собой раздел прикладной математики, ориентированный на разработку методов сбора, систематизации, анализа и интерпретации данных наблюдений с целью выявления статистических закономерностей и построения обоснованных выводов о свойствах генеральной совокупности.

Основные задачи математической статистики:

- оценка неизвестных параметров распределений (точечные и интервальные оценки);
- проверка статистических гипотез о параметрах или виде распределения;
- построение регрессионных моделей и анализ зависимостей между переменными;
- кластерный и факторный анализ для выявления структуры данных;
- планирование эксперимента и оптимизация процедур сбора данных.

Ключевым понятием является **генеральная совокупность** — множество всех объектов, обладающих интересующим исследователя признаком. Непосредственное изучение всей генеральной совокупности зачастую невозможно из-за её объёма, стоимости или разрушающего характера измерений, поэтому используют **выборку** — подмножество элементов, отобранных из генеральной совокупности для анализа.

Виды выборки:

- **Простая случайная выборка**: каждый элемент генеральной совокупности имеет равную вероятность попасть в выборку. Реализуется через случайный отбор (например, с помощью генератора случайных чисел).
- **Стратифицированная выборка**: генеральная совокупность делится на непересекающиеся подгруппы (страты) по существенному признаку, затем из каждой страты отбирается простая случайная выборка. Применяется для повышения точности оценок при наличии выраженной гетерогенности совокупности.
- **Кластерная выборка**: совокупность разбивается на группы (кластеры), случайным образом отбираются некоторые кластеры, и далее исследуются все элементы выбранных кластеров либо проводится внутрикластерная выборка. Эффективна при территориальной рассредоточенности объектов.
- **Систематическая выборка**: элементы упорядочиваются, выбирается начальная точка случайным образом, а затем отбираются элементы через фиксированный интервал.
- **Неслучайные выборки** (удобная, квотная и др.) — используются в разведочных исследованиях, но не позволяют корректно применять вероятностные методы оценивания и проверки гипотез.

Методы математической статистики подразделяются на:

- описательные (дескриптивные) — направлены на компактное представление данных (таблицы, графики, числовые характеристики);
- индуктивные (выводные) — позволяют делать обобщения о генеральной совокупности на основе выборочных данных (оценивание, проверка гипотез).

2. Числовые характеристики вариационного ряда

Вариационный ряд — упорядоченная по возрастанию последовательность значений выборки $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Он служит основой для построения эмпирических функций распределения и расчёта выборочных характеристик.

Основные числовые характеристики

1. **Выборочное среднее** $\bar{x}$:

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

   Оценивает математическое ожидание генеральной совокупности, является несмещённой и состоятельной оценкой.

2. **Выборочная дисперсия** s^2:

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

   Использование делителя $n - 1$ (а не n) обеспечивает несмещённость оценки дисперсии.

3. **Выборочное стандартное отклонение** s:

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

   Характеризует средний разброс значений относительно выборочного среднего.

4. **Медиана** Me:
   - при нечётном n: $Me = x_{\left(\frac{n+1}{2}\right)}$;
   - при чётном n: $Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$.
     Медиана устойчива к выбросам и является робастной характеристикой центра распределения.
5. **Мода** Mo — значение, встречающееся в выборке с наибольшей частотой. В непрерывных распределениях может определяться по гистограмме как центр интервала с максимальной частотой.
6. **Квартили и межквартильный размах**:
   - Q_1 (первый квартиль) — 25-й процентиль;
   - Q_3 (третий квартиль) — 75-й процентиль;
   - межквартильный размах $IQR = Q_3 - Q_1$ — мера разброса, устойчивая к выбросам.
7. **Коэффициент асимметрии** (скошенности) и **эксцесс** — характеризуют форму распределения. Выборочные аналоги строятся через центральные моменты 3-го и 4-го порядков.

Алгоритмы действий при анализе вариационного ряда

Алгоритм расчёта основных характеристик:

1. Упорядочить выборку по возрастанию — получить вариационный ряд.
2. Вычислить выборочное среднее $\bar{x}$.
3. Рассчитать выборочную дисперсию s^2 и стандартное отклонение s.
4. Определить медиану Me по правилам для чётного/нечётного n.
5. Найти моду Mo (по частотам или по гистограмме).
6. Вычислить квартили Q_1, Q_3 и межквартильный размах IQR.
7. При необходимости — оценить асимметрию и эксцесс для анализа формы распределения.
8. Построить описательную статистику и визуализировать данные (гистограмма, boxplot).

Алгоритм проверки репрезентативности выборки:

1. Убедиться, что метод отбора соответствует целям исследования (случайный, стратифицированный и т. д.).
2. Проверить наличие систематических смещений (например, пропуски данных, неравные вероятности отбора).
3. Сравнить выборочные характеристики (среднее, дисперсию) с доступными данными о генеральной совокупности (если таковые имеются).
4. Оценить объём выборки: для нормального распределения часто используется правило «не менее 30 наблюдений» для применения центральной предельной теоремы; для малых выборок или сильно асимметричных распределений требуется более осторожный подход.
5. Провести анализ на наличие выбросов (например, по правилу $1{,}5 \cdot IQR$) и оценить их влияние на выводы.