Начальный анализ статистической информации на основе группировки данных

Структура лекции

1. Основные понятия прикладной статистики
2. Переменные и наблюдения. Типы переменных
3. Группировка данных. Ряд распределения. Таблицы частот

Основные понятия прикладной статистики 3

Цель - определить основные понятия теории вероятностей и статистики, на которые опирается анализ данных изменчивой (случайной) природы.

Статистика изучает числа, чтобы обнаружить в них закономерности.

Явления (ситуации), в которых результат полностью определяется влияющими на него факторами, называются *детерминированными* или *закономерными,* а те, в которых это не выполняется — *недетерминированными* или *стохастическими.*

Для описания явлений с неопределенным исходом (как в повседневной жизни, так и в науке) используется *идея случайности:*

Методы *математической статистики* позволяют оценивать параметры имеющихся закономерностей, проверять те или иные гипотезы об этих закономерностях и т.д.

Основные понятия прикладной статистики

- *События и их вероятности P(A)*
- *Измерение вероятности*
- *Случайные величины. Функции распределения*
- *Числовые характеристики распределения вероятностей*
- *Независимые и зависимые случайные величины*
- *Случайный выбор*
- *Выборки и их описание*
- *Ранги и ранжирование*
- *Методы описательной статистики*
- *Наглядные методы описательной статистики*
- *Методы описательной статистики в ППП*

Переменные и наблюдения. Типы переменных

Показатели, описывающие некоторое явление -

(переменные (variables)).

Каждое значение переменной, полученное в результате наблюдения или эксперимента называется **наблюдением (case) или статистическими данными.**

Переменные бывают нескольких **типов**: **номинальные (категориальные), порядковые (ординальные, ранговые), интервальные**.

Типы статистических данных

Количественные данные отражают в единой шкале измерений некоторый признак (объем продаж, операционные расходы, число посетителей торгового центра и т.д.).

Делят на **дискретные** количественные данные и **непрерывные**.

Ряд данных может иметь **качественный** характер (иногда им присваивают определенные числовые значения).

ТРЕБОВАНИЯ, ПРЕДЪЯВЛЯЕМЫЕ К СТАТИСТИЧЕСКОЙ ВЫБОРКЕ 9

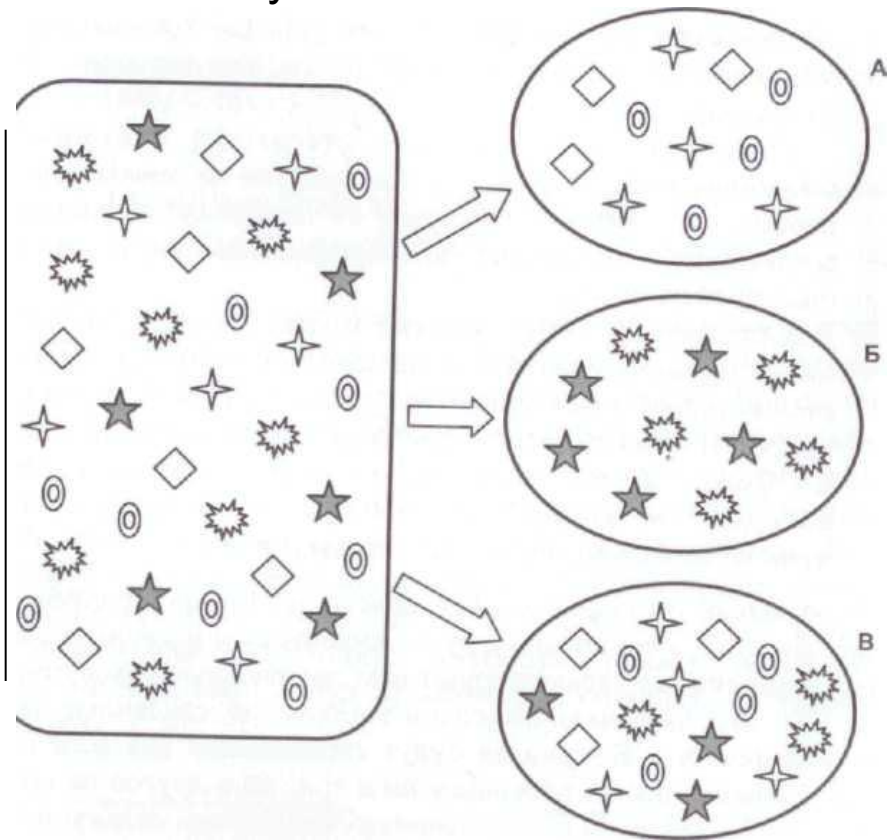
Генеральная совокупность содержит все элементы или все данные, соответствующие изучаемому объекту или явлению.

Выборка – часть данных из генеральной совокупности.

ТРЕБОВАНИЯ, ПРЕДЪЯВЛЯЕМЫЕ К СТАТИСТИЧЕСКОЙ ВЫБОРКЕ 10

Основная цель формирования выборки — *эффективное использование* ее состава в качестве исходной информации для получения правдоподобных (достоверных) выводов обо всех объектах генеральной совокупности;

Основное требование при формировании выборки — **репрезентативность** (представительность). Выборка должна в максимальной степени (как в «капле воды») отражать свойства, структуру генеральной совокупности и ее объектов.



Число элементов выборки (N) должно составлять не менее 10% объема генеральной совокупности. При этом крайне желательно, чтобы общее число элементов (число наблюдений) в выборке было не менее 30 ($N \geq 30$).

Обработка и анализ статистической информации 12

В практических задачах имеем совокупность наблюдений $x_1, x_2, \ldots, x_n$ на основе которых требуется сделать те или иные выводы.

Возникает задача компактного описания имеющихся наблюдений

Идеальное описание такое: в виде утверждения,

Что $x_1, x_2, \ldots, x_n$

являются выборкой, т.е. независимыми реализациями случайной величины ξ с известным законом распределения *F(x).*

Это позволило бы теоретически провести расчеты всех необходимых исследователю характеристик наблюдаемого явления.

Обработка и анализ статистической информации

Определение. ***Методами описательной статистики*** *принято называть методы описания выборок с помощью различных показателей и графиков.*

1. Показатели положения описывают положение данных на числовой оси. Примеры таких показателей — минимальный и максимальный элементы выборки (первый и последний члены вариационного ряда), верхний и нижний квартили (они ограничивают зону, в которую попадают 50% центральных элементов выборки). Наконец, сведения о середине совокупности могут дать выборочное среднее значение, выборочная медиана и другие аналогичные характеристики.

Обработка и анализ статистической информации 15

2. Показатели разброса описывают степень разброса данных относительно своего центра. К ним в первую очередь относятся: дисперсия выборки, стандартное отклонение, размах выборки (разность между максимальным и минимальным элементами), межквартильный размах (разность между верхней и нижней квартилью), коэффициент эксцесса и т.п.

Обработка и анализ статистической информации

3. Показатели асимметрии: отвечает на вопрос о симметрии распределения данных около своего центра. К ней можно отнести: коэффициент асимметрии, положение выборочной медианы относительно выборочного среднего и относительно выборочных квартилей, гистограмму и т.д.

Обработка и анализ статистической информации

4. *Показатели, описывающие закон распределения:* дает представление собственно о законе распределения данных. Сюда относятся графики гистограммы и эмпирической функции распределения, таблицы частот.

Наглядные методы описательной статистики

1. Группировка
2. Точечная диаграмма
3. Гистограмма

Наглядные методы описательной статистики 18

Точечная диаграмма

Точечная диаграмма: табличные данные отмечаются точками на числовой шкале. Если некоторое число встречается в таблице несколько раз, его представляют соответствующим количеством точек.

Наглядные методы описательной статистики 19

Начальная обработка статистических данных

Группировка данных

— разбиение всего диапазона изменения показателя на группы (интервалы) с подсчетом числа наблюдений (частот), попавших в ту либо иную группу, или их доли (относительных частот). Это позволяет оценить, в каких интервалах значений исследуемая величина появляется чаще, а в каких реже.

1. Находят минимальное *Ymin* и максимальное У*мах* значения среди выборочных данных.
2. Весь диапазон изменения величины *Y* — от *Ymin* до У*мах* — разбивают на *интервалы (карманы)* одинаковой **длины.** Количество интервалов *(k)* и их длину определяют, исходя из содержательного смысла анализируемого показателя и задач исследования. На практике число интервалов обычно выбирают не менее 5 и не более 15.
3. Подсчитывают, сколько наблюдений попало в каждый из таких интервалов, т.е. частоты:
4. Также вычисляют относительные частоты — доли наблюдений, оказавшихся в том или ином интервале, удобнее вычислять в процентах:
5. Результаты вычислений сводят в таблицу (сл. Слайд).
6. В зависимости от цели анализа на основе данных 2-й или 3-й графы таблицы строят график — гистограмму, характеризующую особенности распределения исследуемого показателя в зависимости от его значений.

Начальная обработка статистических данных

Таблица

Характеристика сгруппированных данных

Интервал	Частота n_i число наблюдений, попавших в интервал	Относительная частота P_i *(доля наблюдений, оказавшихся в интервале)*	Относительная частота P_i % *(доля в процентах)*

Наглядные методы описательной статистики 22

Гистограмма

Более наглядное описание данных достигается путем группировки наблюдений в классы. Под группировкой, или классификацией будем понимать некоторое разбиение интервала, содержащего все *n* наблюденных результатов $x_1,...,x_n$ на *m* интервалов, которые будем называть *интервалами группировки.*

Длины интервалов обозначим через $\Delta_1,..., \Delta_m$, а середины интервалов группировки — через $t_1,\ldots, t_m$

Число наблюдений n_{ij} в *j-м* интервале группировки равно количеству x_i, i = 1,...,n, удовлетворяющих неравенству $\left|x_i - t_j\right| < \frac{1}{2}\Delta_j$

Определим величину $h_j = n_j / n$

которая означает частоту попадания наблюдений в *j*-й интервал группировки.

Для того чтобы избавиться от влияния размера интервала группировки на h_j, вводится величина $f_j = h_j / \Delta_j$

Наглядные методы описательной статистики 23

Гистограмма

Определение. *Графическое изображение зависимости частоты попадания элементов выборки от соответствующего интервала группировки называется гистограммой выборки.*

В качестве ординаты здесь берется не сама частота, а частота, деленная на длину интервала группировки. Если все интервалы группировки имеют одинаковую длину, деление на Δ обычно опускают и n_j или h_j используют как ординаты.