Лекция 1

1. *Предмет, методы и задачи статистики*
2. *История возникновения статистики как науки*
3. *Статистическая методология*
4. Этапы статистического исследования
5. Организация статистики

1Определение статистики

Термин "статистика" происходит от латинского слова ***status*** (положение, состояние вещей), что в сочетании ***status quo*** означает политическое состояние государства.

В науку этот термин введен немецким ученым *Готфридом Ахенвалем* (1719 – 1772), и означал он тогда *государствоведение*

1.Определение статистики

В русском языке слово «статистика» используется в нескольких значениях. Существует около тысячи определений статистики.

Для нас важными являются три понятия статистики

Статистикой называют числовые данные, ряды цифр, характеризующие различные стороны жизни государства: политические отношения, экономику, культуру, население, отрасли производства и т.д. – все, что можно выразить цифрами

Некоторые оперативные данные можно получить на сайте ФСГС, которую многие по привычке ошибочно называют Госкомстатом.

Адрес сайта http://www.gks.ru/ или http://www.fsgs.ru/ .

*Статистика – это род практической деятельности людей, цель которой – сбор, обработка и анализ информации, **массовых** данных о тех или иных явлениях*

Третье значение слова «статистика»

Статистика – общетеоретическая наука, разрабатывающая статистическую методологию, т.е. набор приемов, способов сбора, обработки и анализа информации.

Именно к этому понятию статистики ближе всего находится наш курс ОТС

2. История возникновения статистики как науки

Современное определение статистики

Статистика – это общественная наука, которая изучает количественную сторону качественно определённых массовых социально-экономических явлений и процессов, их структуру и распределение, размещение в пространстве, движение во времени, выявляя действующие количественные взаимозависимости, тенденции и закономерности в конкретных условиях места и времени

Специфические особенности статистики как науки

1. Статистика изучает количественную *сторону* явлений, т.е. имеет дело с цифрами. Нас интересует экономическая компонента, которая характеризуется цифрами. Существуют явления, которые, на первый взгляд, не характеризуются цифрами. Однако при более глубоком рассмотрении можно найти цифровые параметры, скажем, качества продукции

2. Статистика исследует не единичные факты, а **массовые** явления. Если событие случилось один раз, то это дело репортеров, не статистики. В

Массовые явления и процессы выступают как множества отдельных фактов, обладающих как индивидуальными, так и общими признаками. Объект статистического исследования называют **статистической совокупностью**

Статистическая совокупность -

это множество единиц, обладающих массовостью, однородностью, определённой целостностью, взаимозависимостью состояний отдельных единиц и наличием вариации

Каждый отдельно взятый элемент данного множества называется единицей статистической совокупности

Единицы совокупности

характеризуются общими свойствами, т.е. *признаками*. Признак – это качество, свойство, типичность всех единиц совокупности

Каждая единица совокупности обладает индивидуальными особенностями и различиями, отличающими их друг от друга, т.е. существует так называемая ***вариация*** (колеблемость) признака.

Вариация – это количественные изменения изучаемого признака от одной единицы совокупности к другой

Взаимосвязь курса статистики с другими дисциплинами

Статистическая закономерность устанавливается на основе анализа массовых данных, это обусловливает ее взаимосвязь с **законом больших чисел**

Этапы статистического исследования

Статистическое исследование

складывается из трех основных этапов

- *Статистическое наблюдение* – процесс сбора статистической информации об общественно-экономических явлениях
- *Сводка и группировка* статистического материала, подсчет итогов, расчет обобщающих показателей, изложение результатов в виде таблиц и графиков
- *Анализ* итоговых показателей, *формулировка выводов* и предложений

Статистическое наблюдение

СН – это

научно организованный по единой программе сбор данных (фактов) о социально-экономических, демографических и других явлениях и процессах общественной жизни в государстве с регистрацией их наиболее существенных признаков в учетной документации для ее последующей сводки, обработки и анализа

Отличительные черты СН

- целенаправленность
- организованность
- массовость
- системность (комплексность)
- сопоставимость
- документированность
- контролируемость
- практичность

Способы получения данных

Документальный

Опросный:

- экспедиционный
- саморегистрационный
- корреспондентский

- анкетный
- явочный

Непосредственный

Сводка и группировка статистических данных

Статистическая сводка – это второй этап статистического исследования. В результате проведения статистического наблюдения получают первичную информацию, характеризующую отдельные единицы изучаемой совокупности

✓ **Дальнейшая задача статистики: первичные данные систематизируются и с помощью обобщающих показателей дается сводная характеристика всей совокупности**

Сводка состоит из следующих этапов:

1. **Выбор группировочного признака**

2. **Определение порядка формирования групп**

3. **Разработка системы статистических показателей для характеристики отдельных групп и совокупности в целом**

4. **Разработка макетов статистических таблиц для представления результатов сводки**

✓ **Группировка представляет собой метод, при котором вся исследуемая совокупность разделяется на группы по какому-то существенному признаку**

Основание группировки

✓**Признак, лежащий в основе группировки, называется группировочным или основанием группировки**

✓**В зависимости от вида группировочных признаков различают группировки по количественным и качественным (атрибутивным) признакам**

➢**Группировка по одному признаку называется простой. Группировка по двум и более признакам называется сложной (комбинационной)**

3. Определение количества выделяемых групп

Важнейшим вопросом является определение количества выделяемых групп.

Если в основании группировки лежит атрибутивный признак, то количество выделяемых групп определяется самим этим признаком. Например, производя группировку студентов, посещающих бассейн, по полу, выделяют две группы: мужчин и женщин

Если в основании группировки лежит количественный признак, то производят специальные расчеты для определения количества групп и величин интервалов группировки

Интервал группировки -

это количественное значение, которое определяется как разность между максимальным и минимальным значениями признака в каждой группе

Интервалы группировки могут быть:

Равные

Неравные

Группировки с равными интервалами

применяются в тех случаях, когда вариация признака проявляется в сравнительно узких границах и распределение является практически равномерным

Для группировок с равными интервалами величина интервала h определяется как отношение разности между максимальным и минимальным значениями признака к количеству выделяемых групп:

$$h = \frac{x_{max} - x_{min}}{n} = \frac{R}{n},$$

R – разность между максимальным и минимальным значениями признака (размах вариации)

Xmax – максимальное значение признака в совокупности

Xmin – минимальное значение признака в совокупности

n – количество выделяемых групп

✓ Оптимальное количество групп определяется по формуле Стерджеса:

$$n = 1 + 3{,}322 \lg N,$$

n – количество образуемых групп;

N – число единиц совокупности

- **Интервалы групп могут быть открытые (указана одна из границ) и закрытые (указаны и верхняя и нижняя граница интервала). Величина открытого интервала приравнивается к величине смежного с ним интервала**

- **После определения группировочного признака, количества групп и интервалов группировки данные сводки и группировки представляются в виде рядов распределения и оформляются в виде таблиц**

Статистический ряд распределения представляет собой упорядоченное распределение единиц изучаемой совокупности на группы по определенному варьирующему признаку

Виды рядов распределения

В зависимости от признака, положенного в основу образования ряда распределения, различают:

атрибутивные
вариационные

Атрибутивными

называют ряды распределения, построенные по качественным признакам.

Примерами атрибутивного распределения может служить распределение населения по полу, национальности, месту проживания

Вариационными

называются ряды распределения, построенные по количественному признаку (в порядке возрастания или убывания признака)

Распределение студентов по возрасту, росту

Вариационный ряд **распределения состоит из двух элементов: вариант и частот.**

Количественные значения признака в вариационном ряду распределения называются вариантами и обозначаются x_i.

Частоты – это числа, показывающие: сколько раз в совокупности встречается данное значение признака, и обозначаются f_i

Сумма всех частот равна численности всей совокупности

Частость –

относительное выражение частоты, представляет собой отношение частоты к сумме частот, обозначается p_i.

Может выражаться в процентах:

$$p_i = \frac{f_i}{\sum f_i} \cdot 100\%$$

Сумма всех частостей, выраженных в процентах, равна 100 %, в долях – 1

- ✓ **В зависимости от характера вариации признака вариационные ряды распределения подразделяются на дискретные и интервальные**

- ✓ **Если варианты признаков представлены в виде целых чисел (например, число детей в семье), то такой вариационный ряд называется *дискретным***

- ✓ **Когда значения признака выражены в виде интервалов, это интервальный ряд**

- **Вариационные ряды распределения представляют в виде таблицы, состоящей из двух колонок. В первой колонке приводятся отдельные значения варьирующего признака, т.е. варианты. Во второй – числа, показывающие, сколько раз в совокупности встречается данная варианта, т.е. частоты**

Накопленная (кумулятивная) частота –

Показывает, какое число единиц совокупности имеет величину варианты не большую данной:

$$S_{i+1} = S_i + f_{i+1},$$

где **S** – накопленная частота,

f – частота

Накопленная частость –

рассчитывается аналогично накопленной частоте.

Плотность распределения вариационного ряда:

- абсолютная;
- относительная

Относительная плотность распределения вариационного ряда

Показывает долю единиц совокупности, приходящуюся на единицу величины интервала:

$$\Pi_i^o = \frac{P_i}{h_i}$$

Абсолютная плотность распределения вариационного ряда

Показывает сколько единиц совокупности приходится на одну единицу величины интервала:

$$\Pi_i^a = \frac{f_i}{h_i}$$

- **Для графического изображения дискретного вариационного ряда применяется полигон распределения**

Гистограмма

Применяется для изображения только интервальных вариационных рядов.

При этом по оси абсцисс откладываются интервалы, а по оси ординат – частоты или частости в случае равенства интервалов, или плотности распределения в случае неравенства интервалов

- **Любую гистограмму можно преобразовать в полигон распределения. Для этого достаточно последовательно соединить середины верхних оснований образованных прямоугольников**

В ряде случаев для графического изображения интервальных вариационных рядов применяется кумулята.
Для ее построения сначала необходимо рассчитать накопленные частоты. Они определяются путем последовательного суммирования частот предшествующих интервалов и обозначаются *S*.

Накопленные частоты показывают, сколько единиц совокупности имеют значение признака не больше, чем рассматриваемое.

Накопленная частота последнего интервала должна быть равна сумме частот, т.е. численности единиц совокупности. При построении кумуляты нижней границе первого интервала присваивается накопленная частота, равная 0, и вся накопленная частота интервала присваивается его верхней границе. Для построения кумуляты на оси абцисс

откладывают отрезки, соответствующие интервалам значений признака, на оси ординат – накопленные частота

- На практике приходится пользоваться уже имеющимися группировками, которые могут быть несопоставимы из-за неодинаковых границ интервалов или различного количества выделяемых групп. Для приведения таких группировок к сопоставимому виду используется метод вторичной группировки
- **Вторичная группировка – это образование новых групп на основе ранее произведенной группировки. Применяют два способа образования новых групп на основе ранее произведенной группировки**

Первый способ

состоит в укрупнении первоначальных интервалов. Это наиболее простой и распространенный способ

Второй способ

называется методом долевой перегруппировки и состоит в том, что за каждой группой закрепляется определенная доля единиц совокупности.

Абсолютные и относительные статистические величины

1.Абсолютные величины и их виды

- АВ – показатели, выражающие размеры социально-экономических явлений числом единиц или величиной характеризующих их признаков в данных условиях места и времени

- АВ – количественные показатели, выражающие общую численность, размеры (объемы, уровни) и другие характеристики изучаемого процесса или явления

Виды абсолютных величин

индивидуальные суммарные

Абсолютные величины

являются основой для расчета разных относительных статистических показателей

- **Относительные величины** в статистике представляют собой частное от деления двух статистических величин и характеризуют количественное соотношение между ними

Важное свойство – относительная величина абстрагирует различия абсолютных величин и позволяет сравнивать такие явления, абсолютные размеры которых непосредственно несопоставимы

Средние величины

Средняя величина – это обобщающая количественная характеристика совокупности по изучаемому признаку в конкретных условиях места и времени.

Средняя величина отражает то общее и типичное, что присуще единицам данной совокупности

В средних величинах погашаются индивидуальные отклонения, соответствующие отдельным единицам совокупности. Чтобы средняя величина имела смысл, она должна рассчитываться для однородной совокупности

Используя среднюю, мы можем одним числом охарактеризовать изучаемое явление. По уточненным данным Всероссийской переписи населения 2002 года, средний размер семьи составляет 2,7 чел. В городских населенных пунктах – 2,7. В сельских – 2,8.

Необходимые условия для расчета СВ – качественная однородность совокупности: все единицы совокупности должны обладать изучаемым признаком.

Виды средних величин

средняя арифметическая:

$$\overline{x} = \frac{\sum x}{n};$$

средняя квадратическая:

$$\overline{x}_q = \sqrt{\frac{\sum x^2}{n}};$$

средняя геометрическая:

$$\overline{x_g} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i};$$

средняя гармоническая:

$$\overline{x}_h = \frac{n}{\sum \frac{1}{\overline{x}}}$$

Правило мажорантности

$$\overline{x}_q \rangle \overline{x} \rangle \overline{x}_g \rangle \overline{\overline{x}}_h$$

Существуют две формулы средней арифметической:

$$\bar{x} = \frac{\sum x}{n} - простая;$$

$$\bar{x} = \frac{\sum x \cdot f}{\sum f} - взвешенная,$$

где *f* - веса

- Средняя арифметическая простая применяется, когда есть перечисление вариант и нет никаких группировок.

В числителе мы собираем сумму вариант, в знаменателе – количество вариант

- Средняя арифметическая взвешенная используется при появлении группировок. Это самая распространенная степенная средняя

- Если f – частость (дается удельный вес в совокупности), то классическая формула средней арифметической взвешенной не применяется, используют ее модификацию:

$$\overline{x_i} = \sum_{x=1}^{n} x_i * d_i,$$

$$d_i = \frac{f_i}{\sum f_i};\ \sum f_i - численность\_совокупности$$

$$f_i - число\_единиц\_с\_определенным\ значением\_варианты;$$

Свойства средней арифметической

1. Произведение средней арифметической и суммы частот равно общему объему изучаемого события в совокупности :

$$\bar{x} * \sum f_i = \sum x_i f_i$$

2. Сумма отклонений всех вариант от средней величины всегда равна 0:

Это значит, что в средней

арифметической
взаимопогашаются
отклонения от средней

3. Если каждую варианту уменьшить на постоянную величину a, расчет средней возможен, но полученная средняя будет меньше на a:

4. Если все варианты уменьшить в одно и то же число раз, то средняя арифметическая уменьшится в то же число раз:

5. Если все веса разделить на какую-либо константу a, то новая средняя от этого не изменится:

5. При расчете средней весовой показатель берется на том же уровне и в числителе, и в знаменателе

Средняя гармоническая

- СГ- это обратная величина средней арифметической. Бывает простая и взвешенная СГ. Чаще используется взвешенная формула

Существуют две формулы для расчета средней гармонической величины:

$$\overline{x}_h = \frac{n}{\sum \frac{1}{x}} - \textit{простая},$$

$$\overline{x}_h = \frac{\sum W}{\sum \frac{W}{x}} - \textit{взвешенная}.$$

где *W*- сложный вес, объем события по группе, по конкретному значению

Сложный (мнимый) вес: $W_i = x_i * f_i$

Средняя гармоническая применяется в том случае, когда в качестве весов выступают объемы изучаемого признака.

Иногда возникает проблема: какую формулу использовать – среднюю гармоническую или среднюю арифметическую? Подходит та формула, у которой и в числителе и знаменателе будут величины, обладающие смыслом

Арифметическая или гармоническая?

- Подсказка:
- Если по исходной информации дается осредняемая величина (варианта) и знаменатель логической формулы, то используется САВ.
- Если дается варианта и числитель логической формулы, то используется СГВ

Средняя хронологическая

- Эта формула средней применяется для ряда моментных показателей

$$\overline{X} = \frac{\frac{1}{2}x_1 + x_2 + x_3 + \ldots + x_{n-1} + \frac{1}{2}x_n}{n-1}$$

- Если необходимо подсчитать среднюю для двух моментных показателей, то формула средней хронологической превращается в формулу средней арифметической простой

$$\overline{X} = \frac{\frac{1}{2}x_1 + \frac{1}{2}x_2}{1} = \frac{x_1 + x_2}{2}$$

Структурные средние

Структурные средние применяются для первоначального анализа распределения признаков в совокупности

Мода – это значение признака, встречающееся в совокупности наибольшее число раз.

В быту слово «мода» фактически имеет обратный смысл

Мода – это наиболее часто встречающаяся варианта вариационного ряда.

Для дискретного ряда это та варианта, которой соответствует наибольшая частота

Для интервального ряда с равными интервалами мода определяется при помощи следующей формулы:

$$M_o = x_{M_o} + h_{M_o} \cdot \frac{f_2 - f_1}{2f_2 - f_1 - f_3}$$

где x_{M_o} - начало модального интервала;

h_{Mo} - величина модального интервала;

f_2 - частота модального интервала;

f_1 - частота предмодального интервала;

f_3 - частота послемодального интервала

- Если модальный интервал первый или последний, то недостающая частота (предмодальная или послемодальная) берется равной нулю

  - В интервальном ряду как по формуле, так и графически мода вычисляется точнее

- Для определения моды дискретного ряда строится полигон распределения. Расстояние от оси ординат до наивысшей точки графика есть мода

- Если в дискретном ряду несколько вариант имеют наибольшую частоту (что встречается достаточно редко), то мода определяется как средняя арифметическая из всех модальных вариант

Медиана

- Это центральное, серединное значение ряда.

  Обозначение: Me

Me - значение признака у единицы, находящейся в середине ранжированной (упорядоченной) совокупности

Это варианта, лежащая в середине вариационного ряда и делящая его на две равные части.

В дискретном ряду Ме находится по определению, а в интервальном ряду – по формуле

- Если дискретный ряд содержит нечетное количество вариант, то находится та единственная варианта, справа и слева от которой находится одинаковое число вариант:

$$Me = x_{\frac{n+1}{2}}$$

- Если дискретный ряд содержит четное количество вариант, то находятся две варианты, справа и слева от которых располагается одинаковое количество вариант. Ме равна средней арифметической из двух значений:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2}$$

Для дискретного ряда медианой является та варианта, для которой накопленная частота впервые превышает половину от суммы частот

Для интервального ряда медиана определяется по следующей формуле:

$$Me = x_{Me} + h_{Me} \cdot \frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}},$$

где x_{Me} - начало медианного интервала;

h_{Me} - величина медианного интервала;

f_{Me} - частота медианного интервала;

S_{Me-1} - накопленная частота предмедианного интервала

Для графического определения медианы последнюю ординату кумуляты делят пополам. Через полученную точку проводят прямую, параллельную оси *X* до пересечения ее с кумулятой. Абсцисса точки пересечения является медианой

- В практических расчетах Мо и Ме могут быть величинами, далеко отстоящими друг от друга. Для более четкой фиксации характера распределения используют другие структурные средние

Квартили

Это варианты, которые делят ранжированную совокупность на четыре равные части:

Q_1 1:3;

Q_2 2:2 (Q_2=Me);

Q_3 3:1

- Первый (нижний) квартиль отсекает от совокупности ¼ часть единиц с минимальными значениями, а третий (верхний) отсекает ¼ часть единиц с максимальными значениями

- Мы как бы отбрасываем нетипичные, случайные значения признака. С помощью квартилей мы определяем границы, где находятся 50% единиц, наиболее характерные для этой совокупности

Для расчета Q_1 (первого квартиля) используется следующая формула:

$$Q_1 = x_{Q_1} + h_{Q_1} \cdot \frac{\frac{\sum f}{4} - S_{Q_1 - 1}}{f_{Q_1}},$$

где x_{Q_1} - начало интервала, содержащего 1-й квартиль;

h_{Q_1} - величина интервала, содержащего 1-й квартиль;

$S_{Q_1 - 1}$ - накопленная частота предшествующего интервала;

f_{Q_1} - частота интервала, содержащего Q_1

Интервалом, содержащим Q_1, является тот интервал, для которого накопленная частота впервые превышает ¼ от суммы частот

Для расчета Q_3 используется формула:

$$Q_3 = x_{Q_3} + h_{Q_3} \cdot \frac{\frac{3\sum f}{4} - S_{Q_3-1}}{f_{Q_3}}.$$

Все обозначения аналогичны Q_1.
Интервалом, содержащим Q_3,
является тот интервал, для которого
накопленная частота впервые превышает ¾
от суммы частот

Децили -

это варианты, которые делят ранжированную совокупность на 10 равных частей

Общая формула для расчета децилей:

$$D_i = x_{D_i} + h_{Di} \cdot \frac{i \cdot \frac{\sum f}{10} - S_{Di-1}}{f_{Di}},$$

где x_{Di} - начало интервала,содержащего *i*-й дециль;

h_{Di} - величина интервала, содержащего *i*-й дециль;

f_{Di} - частота интервала, содержащего D_i;

S_{Di-1} - накопленная частота предшествующего интервала

Интервалом, содержащим D_i ,является тот интервал, для которого накопленная

частота впервые превышает i/10 от суммы частот

Перцентиль

- П делит ранжированную совокупность на 100 равных частей. Формулы аналогичны формулам медианы, квартиля и дециля