Лекция 2

МЕТОДЫ ИССЛЕДОВАНИЙ В СТАТИСТИКЕ

Статистическое исследование можно условно разделить на три этапа - сбор данных, их первичная обработка и статистический анализ. Первый и второй этапы тесно связаны с той областью знаний, в которой проводится статистическое исследование, поэтому являются предметом детального изучения в отраслевых статистиках. В разделе "Описательная статистика" будут рассмотрены лишь основные понятия и подходы, применяемые на этих этапах. С другой стороны, рассматриваемые методы *анализа статистических данных* являются достаточно универсальными и не зависящими от природы самих данных.

Описательная статистика

Статистическое наблюдение

Статистическое наблюдение - это научно организованный сбор массовых данных об исследуемых процессах или явлениях. Статистические данные представляют собой, образно говоря, "сырье", полученное в результате статистического наблюдения, которое в процессе обработки становится информацией. Результаты статистического исследования будут ценны лишь в том случае, если они базируются на конкретном материале. Применение сложных методов анализа может оказаться неэффективным, если нет уверенности в первичном материале. Важнейшими требованиями к статистическим данным является их достоверность, полнота и сопоставимость.

Организация наблюдения

Организация статистического наблюдения включает в себя: (i) определение целей и задач; (ii) выбор объекта (статистической совокупности) и единиц, подлежащих обследованию; (iii) составление программы наблюдения.

Всякое наблюдение должно быть максимально подчинено поставленной *целы и задачам,* поэтому их следует четко сформулировать. Например, всеобщая перепись населения проводится с целью получить сведения о численности населения в отдельных

регионах и населенных пунктах страны, о его половозрастной структуре, занятости, условиях проживания, составе семьи и ряде других характеристик.

Одновременно с определением целей устанавливается объект статистического наблюдения, *статистическая совокупность,* состоящая из *единиц* наблюдения. Например, в случае всеобщей переписи населения статистической совокупностью является все население страны, а единицами наблюдения - отдельные граждане. В случае обследования фермерских хозяйств совокупностью являются все хозяйства, например, данной области, а единицами - отдельные хозяйства.

Единицы наблюдения обладают множеством *признаков,* то есть качествами, свойствами, по которым устанавливается их сходство и различия. Учесть все признаки в ходе обследования невозможно, и, главное, в этом нет необходимости. Учитываемые признаки выбираются исходя из целей и задач исследования. Например, для исследования фермерских хозяйств, признаками является: количество членов семьи, размер участка, количество техники, размер урожая, полученный доход, взятые кредиты.

Существенным для дальнейшего рассмотрения является деление признаков на атрибутивные и количественные.

Атрибутивные признаки отражают качественное состояние наблюдаемого объекта и выражаются содержательными понятиями (пол, образование, профессия, вероисповедание, вид экономической деятельности).

Количественные признаки характеризуют размер, объем, величину данного явления и выражаются числами (возраст, стаж работы, размер стипендии, продолжительность рабочей недели, объем продаж и т.п.).

Ошибки статистического наблюдения

Как бы тщательно не был продуман план статистического исследования, как бы точно не выполняли инструкции участвующие в исследовании лица, при любом наблюдении возникают ошибки. Все ошибки можно разделить на преднамеренные и непреднамеренные; непреднамеренные, в свою очередь, могут быть случайными и систематическими.

Непреднамеренные *случайные* ошибки возникают в результате неправильной

работы прибора, скачков напряжения, описок, оговорок, незнания и т.п. как по вине исполнителя (отвечающего), так и по вине регистратора. Как правило, эти ошибки не опасны, так как при большом числе наблюдений они взаимопогашаются, нейтрализуются.

Непреднамеренные *систематические* ошибки возникают при опросе за счет округлений количественных показателей (неправильной калибровки, возраста, стажа работы, дохода) или за счет неточности измерительных приборов. Замечено, например, что при регистрации возраста путем опроса возраст наиболее часто округляется до чисел, оканчивающихся на 0 или 5. В результате оказывается, что 40-летних по записям значительно больше, чем 39- и 41-летних. Это явление получило в статистике название аккумуляция возрастов. Такие погрешности приходится исправлять уже при обработке собранных данных.

Преднамеренные ошибки, как это ясно уже из названия, возникают в силу сознательного стремления лиц, дающих сведения, исказить истину - уменьшить или увеличить значение признака. Очевидно, что преднамеренные ошибки содержат искажения в одном направлении, поэтому этот вид ошибок наиболее опасен для статистического исследования, их необходимо выявлять и устранять.

Все рассмотренные виды ошибок называют ошибками регистрации. Кроме них в случае, если проводится не сплошное, а *выборочное* наблюдение, возникают *ошибки репрезентативности*.

Группировка статистических данных

В результате проведения статистического наблюдения получают данные о признаках каждой единицы обследованной совокупности. Однако для того чтобы выявить характеристики совокупности в целом, эти данные необходимо обобщить и систематизировать. *Группировкой* называется разделение на группы единиц статистической совокупности, однородных по одному или нескольким признакам. Группировки по одному признаку делятся на несколько типов в зависимости от того, является ли группировочный признак атрибутивным или количественным, и как много различных значений он принимает. Для обозначения различий в значениях признака у единиц совокупности используют термин *вариация признака,* а совокупность значений одного признака на-

зывают *рядом* или *вариационным рядом.*

Группировка по атрибутивному признаку

Рассмотрим основные элементы группировки на следующем примере.

Пример 1. В ходе проведения опроса 25 респондентов - мужчин получена следующая информация об их семейном положении (Ж - женат, Х - холост, Р - разведен, В - вдовец):

Ж Х В Р Ж Ж Ж Ж Ж Х

Х Х Р Ж Ж В Ж Ж В Р Р Х Х

Х Р

Выполнить группировку собранных данных.

Решение. Выполним группировку, последовательно заполняя колонки табл. 1.1. Группировочный признак принимает четыре различных значений, поэтому все данные подразделим на четыре группы (первая колонка). Далее, для каждой группы определим ее *частоту* - величину, показывающую, сколько раз признак принял значение данной группы. Удобный способ подсчета основан на использовании вспомогательной колонки (табл. 2, вторая колонка). Последовательно просматриваем *ряд* данных и для каждого значения ставим черточку напротив соответствующей группы. После окончания просмотра частота группы определяется подсчетом стоящих напротив нее черточек. Сумма всех частот равняется общему числу единиц наблюдения -- 25.

Таблица 1.1.

Семейное	*Подсчет*	*Частота*	*Частость (%)*
Х	///////	7	28
Ж	//////////	10	40
Р	/////	5	20
В	///	3	12
Всего N		*25*	*100%*

По сгруппированным данным легко можно сказать, например, что среди опрошенных наибольшее число мужчин женаты (10), холостых несколько меньше (7).

Полученная таблица - совокупность значений группировочного признака и их частот имеет специальное название - *частотное распределение.* В случае, когда группировка проводится по атрибутивному признаку, частотное распределение называется *атрибутивным.*

Кроме частот, для группы рассчитывают еще одну величину, *частость,* равную отношению частоты к общему количеству единиц наблюдения и выраженную в процентах: $w_i = f_i / N \times 100\%$. Частость показывает, какая доля всех единиц попала в данную группу. По смыслу частость совпадает с уже рассмотренной ранее относительной частотой, однако в российской статистической литературе чаще используется термин частость.

Группировка по количественному признаку с малым количеством значений

В случае, когда рассматриваются результаты наблюдений по количественному признаку, способ группировки дополнительно зависит от того, является ли признак дискретным или непрерывным, и какое количество различных значений он принимает. В случае дискретного признака с небольшим количеством значений группировка проводится аналогично случаю атрибутивного признака - каждое значение составляет отдельную группу. Соответствующее распределение называют *дискретным частотным распределением.*

Пример 2. Преподаватель провел контрольный тест среди 20 студентов и зафиксировал время (в мин.), затраченное каждым из них на выполнение работы:

```
8   8   9   8   5   9   9     10  7   7
8   7   8   7   6   5   7     10  8   9
```

Выполнить группировку данных (построить частотное распределение).

Решение. Группировочный признак имеет небольшое количество числовых значений - 5, 6, 7, 8, 9, 10, поэтому каждое из них можно поместить в отдельную группу (упорядочив по возрастанию), табл. 1.2, колонка 2. Затем по исходным данным находим частоты, колонка 3. Дискретное частотное распределение построено.

Таблица 1.2.

№	*Время выполнения*	*Частота* f_i	*Частость* w_i *(%)*	*Накопленная частота* cf_i	*Накопленная частость* cw_i
1	5	*2*	10	2	10
2	6	1	5	3	15
3	7	5	25	8	40
4	8	7	35	15	75
5	9	3	15	18	90
6	10	2	10	20	100
	Всего *N*	20	*100%*		

Основная часть студентов (больше половины) затратила на тест 7 или 8 минут, двое студентов выполнили тест существенно быстрее (за 5 минут), но были и двое таких, которым понадобилось 10 минут.

Кроме колонок с частотой и частостью, в таблице присутствуют две новые величины, табл. 1.2, колонки 5, 6. *Накопленная частота* вычисляется как сумма частоты данной группы и накопленной частоты предыдущей группы:

$$cf_i = cf_{i-1} + f_i, \quad i = 2,\ldots,n, \; cf_1 = f_1.$$

Накопленная частота i-й группы показывает, какое количество студентов затратили на выполнение теста не более x_i минут. Например, $cf_4 = 15$ означает, что 15 студентов затратили на тест не более 8 минут. *Накопленная частость* выражает ту же величину в частях от целого; $cw_5 = 90\%$ означает, что 90% студентов на выполнение теста потребовалось не более 9 минут.

Группировка по количественному признаку с большим количеством значений

В случае, когда группировка проводится по количественному признаку с большим количеством значений, в качестве групп применяются не сами значения, а *интервалы* значений. Частота каждого интервала показывает, какое количество значений попало в этот интервал. Получившееся частотное распределение называется *интервальным.* Построение интервального распределения можно разбить на следующие этапы.

Алгоритм построения интервального частотного распределения

1) Выбрать наибольшее $x_{\max}$, наименьшее $x_{\min}$ значения признака и вычислить размах вариации $R = x_{\max} - x_{\min}$.

2) Задать количество интервалов K.

3) Определить ширину интервала $H = R/K$, округлить до числа знаков, равного числу знаков у результатов наблюдения.

4) Рассчитать нижние $x_{1,i}$ и верхние $x_{h,i}$ границы интервалов:

$x_{l,i} = x_{l-1} + H$, $i = 2,\ldots,K$, $x_{l,1} = x_{\min}$;

$x_{h,i} = x_{l,i+1}$, $i = 1,\ldots,K-1$; $x_{h,K} = x_{l,K} + H$.

5) Рассчитать средние значения интервалов $x_{m,i} = \left(x_{l,i} + x_{h,i}\right)/2$, $i = 1,\ldots,K$.

6) Для каждого интервала определить частоту, частость, накопленные частоту и частость.

Пример 3. Компания-производитель провела ресурсные испытания партии новых зимних шин с целью определить пробег (в тысячах километров) до критического износа. Получены следующие значения:

10	29	6	33	14	21	18	35	22	38	31
24	27	19	22	23	26	39	34	27		

Построить интервальное частотное распределение.

Решение. Применим описанный выше алгоритм построения распределения.

1) Наибольший пробег составляет 39 тыс. км., наименьший - б тыс. км., размах вариации равен $R = 39 - 6 = 33$.

2) Возьмем количество групп, равное семи, $K = 7$.

3) Ширина интервала $H = R/K = 33/7 \approx 4.7$. Результаты наблюдений заданы с точностью до целого (до тысячи километров), поэтому ширина класса округляется до 5, $H = 5$.

4) Вычисляем нижние границы интервалов:

$x_{l,1} = x_{\min} = 6$;

$x_{l,2} = x_{l,1} + H = 6 + 5 = 11$;

$x_{l,3} = 11 + 5 = 16$;

$x_{l,4} = 16 + 5 = 21$;

$x_{l,5} = 21 + 5 = 26$;

$x_{l,6} = 26 + 5 = 31$;

$x_{l,7} = 31 + 5 = 36$.

Верхние границы интервалов равны нижним границам интервалов, следующих за ними. Последний интервал сверху ограничивается значением $x_{h,7} = 41$. Результаты расчетов помещены в табл. 1.3, колонка 2.

5) Находим средние значения интервалов $x_{m,1} = (6+11)/2 = 8.5$ и т.д., колонка 3.

6) Просматриваем значения ряда, находим частоту для каждого интервала - количество единиц, значение признака которых попадает в данный интервал, колонка 4.

Рассчитываем частости, накопленные частоты и накопленные частости, колонки 5-8.

Таблица 1.3.

№	*Интервал пробега* $[x_{l,i}, x_{h,i}]$	*Середины интервалов,* $x_{m,i}$	*Частота* f_i	*Частость* w_i	*Накопленная частота* cf_i.	*Накопленная частость* cw_i
1	6-11	8.5	*2*	10	*2*	10
2	11-16	13.5	1	5	3	15
3	16-21	18.5	3	15	6	30
4	21-26	23.5	5	25'	11	55
5	26-31	28.5	4	20	15	75
6	31-36	33.5	3	15	18	90
7	36-41	38.5	2	10	20	100

Замечания.

1) Ключевым вопросом при построение интервального распределения является выбор числа групп. Чем больше групп, тем уже интервал и тем точнее будет распределение. Однако, с другой стороны, большое число групп осложняет

восприятие распределение. Обычно на практике используют 7 - 1 0 групп.

Число интервалов также можно приближенно оценить с помощью формулы Стерджеса

$$K = 1 + 3,32 \times \log_{10} N$$

в которой, результат округляется до целого числа. Во многих случая предпочтительным является использование нечетного числа групп.

2) Интервалы, у которых определены обе границы ("от" и "до") называются закрытыми. Использование открытых интервалов (например, "менее 20" или "40 и более") является хотя и нежелательным, но в большинстве случаев неизбежным, так как все крайние случаи ради компактности приходится сводить в один интервал.
3) При определении границ интервалов по представленному алгоритму верхняя и нижняя граница соседних интервалов совпадают, что приводит к неоднозначности, если значение признака точно совпадает с граничным значением (в примере это значения 21 и 26). В этом случае исследователь должен принять решение и отнести эти значения к одному из интервалов (в нашем случае они относились к нижнему). Проблема может быть решена и по-другому, если уменьшить верхние границы на небольшую величину, (например 0.1). Тогда границы интервалов примут вид 6-10.9, 11-15.9 и т.д.