Содержание

Введение

Курс «Статистика» имеет целью дать студентам представление о содержании статистики как научной дисциплины, познакомить с ее основными понятиями, методологией и методиками расчета важнейших статистических показателей.

Слово «статистика» имеет латинское происхождение (от слова status – состояние). Развитие статистической науки, расширение сферы практической статистической работы привели к изменению содержания самого понятия «статистика». В настоящее время это понятие употребляется в трех значениях:

- под статистикой понимают отрасль практической деятельности, которая имеет своей целью сбор, обработку, анализ и публикацию массовых данных о самых различных явлениях общественной жизни (в этом смысле «статистика» выступает как синоним словосочетания «статистический учет»);
- статистикой называют цифровой материал, служащий для характеристики какой-либо области общественных явлений;
- статистикой называется отрасль знания, особая научная дисциплина и соответственно учебный предмет в вузах.

Как и всякая наука «статистика» имеет свой предмет, которым является количественная характеристика массовых общественных явлений в неразрывной связи и их количественной характеристикой. Свой предмет «статистика» изучает при помощи определенных категорий, т.е. понятий, которые отражают наиболее общие и существенные свойства, признаки, связи и отношения между ними.

В дисциплине «Статистика» излагаются вопросы общей теории статистики: предмет, метод статистики как общественной науки, ее значение; способы получения статистической информации, основы ее научной разработки, правила и приемы сравнения данных; виды средних и их применение в статистико-экономическом анализе; показатели вариации; методы изучения временных рядов; индексный метод анализа; основы регрессионного и корреляционного анализа.

В соответствии с учебным планом и рабочей программой курса «Статистика» студенты выполняют 2 контрольные работы.

Для выполнения **контрольной работы № 1 «Показатели вариации»** следует изучить следующие вопросы:

- сводка и группировка статистических данных;
- средние величины;
- абсолютные и относительные показатели вариации;
- правило сложения дисперсий;
- способ условных моментов.

Для выполнения **контрольной работы № 2 «Статистическое изучение взаимосвязи между признаками»** следует изучить следующие вопросы:

- определение формы и вида связи;
- определение параметров уравнения связи;
- определение степени тесноты и существенности связи.

Вариант контрольных работ определяется последней цифрой студенческого билета.

Контрольная работа № 1 состоит из двух задач (№ 1 и № 2), контрольная работа № 2 – из одной задачи (№ 3).

Работы представляются в печатном виде на листах форматом А4 с подробным пояснением.

I. Исходные данные для выполнения контрольных работ

В результате выборочного обследования 10% рабочих авиаремонтного завода (по состоянию на 1 января текущего года) получены следующие данные:

№№ п\п	Разряд	Производственный стаж, лет	Заработная плата, у.е.
1	2	3	4
Цех № 1			
1	4	5	539
2	1	1	487
3	4	7	554
4	2	2	507
5	1	1	490
6	2	5	519
7	3	8	536
8	5	10	574
9	2	0	481
10	3	7	533
11	2	2	515
12	2	3	524
13	5	5	553
14	1	1	479
15	3	4	509
16	3	8	552
17	2	3	526
18	2	1	495
19	1	0	492
20	4	6	562
21	2	5	516
22	1	0	483
23	4	8	531
24	4	12	548

№№ п\п	Разряд	Производственный стаж, лет	Заработная плата, у.е.
25	2	4	521
26	3	7	529
27	3	6	520
28	2	1	475
29	3	8	525
30	1	0	472
31	4	3	553
32	2	4	518
33	1	0	485
34	2	3	508
35	3	8	507
36	5	17	578
37	2	1	505
38	6	23	600
39	3	4	528
40	3	11	538
Цех № 2			
1	3	5	536
2	2	1	501
3	3	3	517
4	4	15	571
5	2	1	492
6	4	19	562
7	1	0	480
8	3	5	541
9	3	7	535
10	2	1	502
11	3	3	528
12	4	12	565
13	4	2	525
14	5	6	536
15	5	8	574
16	3	3	523
17	6	29	571
18	2	3	498
19	4	13	537
20	3	8	530
21	1	1	494
22	2	0	468
23	4	3	513
24	3	9	547

№№ п\п	Разряд	Производственный стаж, лет	Заработная плата, у.е.
25	6	9	594
26	5	12	588
27	1	2	504
28	3	6	523
29	1	0	460
30	4	14	536
31	2	4	517
32	3	5	535
33	3	0	492
34	4	15	553
35	5	8	573
36	2	1	486
37	4	2	543
38	3	4	522
39	3	7	534
40	4	10	558
41	2	4	506
42	2	4	512
43	3	11	552
44	4	5	527
45	4	7	547
46	5	15	595
47	3	4	514
48	3	8	555
49	3	9	524
50	2	4	505
51	4	11	559
52	1	1	491
53	3	9	534
54	4	10	552
55	3	2	526
56	5	21	597
57	3	8	521
58	2	0	483
59	5	13	575
60	2	2	508

2. Варианты заданий для выполнения контрольных работ

Вариант № 0

1. Построить ряд распределения рабочих каждого цеха и всего завода по квалификации (разрядам). Рассчитать относительные величины структуры, характеризующие состав рабочих по квалификации.

2. Определить дисперсию тарифного разряда в каждом цехе и по заводу в целом; среднюю из цеховых дисперсий; межцеховую дисперсию. Объяснить смысл дисперсий. Используя их, проверить правило сложения дисперсий.

3. Определить количественную связь между признаками:

3.1. С помощью графического метода определить форму связи между разрядом и заработной платой рабочих цеха № 1 с № 1 по № 20 включительно (п=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между разрядом рабочих и их заработной платой. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3. Определить степень тесноты между рассматриваемыми признаками.

Вариант № 1

1. Построить ряд распределения рабочих каждого цеха и всего завода по размеру заработной платы, выделив семь групп с равными интервалами.

2. Определить дисперсию заработной платы рабочих в каждом цехе и по заводу в целом; среднюю из цеховых дисперсий; межцеховую дисперсию. Объяснить смысл дисперсий. Используя их, проверить правило сложения дисперсий.

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между тарифным разрядом и заработной платой рабочих цеха № 1 с № 21 по № 40 включительно (п=20).

2. Вычислить параметры уравнения регрессии, характеризующего зависимость между разрядом рабочих и их заработной платой. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3. Определить степень тесноты между рассматриваемыми признаками

Вариант № 2

1. Построить ряд распределения рабочих завода по общему стажу работы, выделив пять групп со следующими специализированными интервалами: 1) менее года; 2) от 1 до 2 лет; 3) от 3 до 5 лет; 4)от 6 до 10 лет; 5) от 11 лет и выше.

2. Определить дисперсию производственного стажа рабочих в каждом цехе и по заводу в целом; среднюю из цеховых дисперсий; межцеховую дисперсию. Объяснить смысл дисперсий. Используя их, проверить правило сложения дисперсий.

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между производственным стажем и заработной платой рабочих цеха № 1 с № 1 по № 20 включительно (п=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между производственным стажем рабочих и их заработной платой. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3.Определить степень тесноты между рассматриваемыми признаками.

Вариант № 3

1. Произвести комбинационное распределение рабочих каждого цеха и завода в целом по разрядам и размеру месячной заработной платы.

2. Рассчитать дисперсию заработной платы в цехе № 1 и № 2. Определить коэффициенты вариации заработной платы рабочих по цехам. Сделать выводы.

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между тарифным разрядом и заработной платой рабочих цеха № 2 с № 1 по № 20 включительно (п=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между тарифным разрядом и заработной платой рабочих. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3.Определить степень тесноты между рассматриваемыми признаками.

Вариант №4

1. Построить комбинационное распределение рабочих каждого цеха и завода в целом по общему стажу работы и разрядам.

2. Рассчитать дисперсию производственного стажа рабочих в цехе № 1 и № 2. Определить коэффициенты вариации производственного стажа рабочих по цехам. Сделать выводы.

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между тарифным разрядом и заработной платой рабочих цеха № 2 с № 21 по № 40 включительно (n=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между тарифным разрядом и заработной платой рабочих. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3.Определить степень тесноты между рассматриваемыми признаками.

Вариант № 5

1. Построить комбинационное распределение рабочих каждого цеха и завода в целом по общему стажу работы и заработной плате.

2. Рассчитать дисперсию тарифного разряда рабочих в цехах № 1 и № 2. Определить коэффициенты вариации тарифного разряда рабочих по цехам. Сделать выводы.

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между тарифным разрядом и заработной платой рабочих цеха № 2 с № 41 по № 60 включительно (n=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между тарифным разрядом и заработной платой рабочих. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3.Определить степень тесноты между рассматриваемыми признаками

Вариант № 6

1. Построить ряд распределения рабочих каждого цеха и завода в целом по размеру заработной платы, выделив 7 групп с равными интервалами. Определить в целом по заводу моду и медиану заработной платы рабочих.

2. Рассчитать среднюю заработную плату и дисперсию заработной платы рабочих завода обычным способом и способом условных моментов.

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между производственным стажем и заработной платой рабочих цеха № 1 с № 21 по № 40 включительно (n=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между производственным стажем и заработной платой рабочих. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3.Определить степень тесноты между рассматриваемыми признаками

Вариант № 7

1. Построить ряд распределения рабочих завода по разрядам. Определить в целом по заводу моду разряда рабочих.

2. Рассчитать абсолютные и относительные показатели вариации тарифного разряда рабочих по цехам. Сделать выводы.

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между производственным стажем и заработной платой рабочих цеха № 2 с № 1 по № 20 включительно (п=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между производственным стажем и заработной платой рабочих. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3. Определить степень тесноты между рассматриваемыми признаками.

Вариант № 8

1.Построить ряд распределения рабочих завода по заработной плате и определить ее медиану.

2. Рассчитать абсолютные и относительные показатели вариации заработной платы рабочих по цехам. Сделать выводы.

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между производственным стажем и заработной платой рабочих цеха № 2 с № 21 по № 40 включительно (п=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между производственным стажем и заработной платой рабочих. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3. Определить степень тесноты между рассматриваемыми признаками.

Вариант № 9

1. Построить ряд распределения рабочих завода по стажу работы, выделив пять групп со следующими специализированными интервалами: 1) менее года; 2) от 1 до 2 лет; 3) от 3 до 5 лет; 4) от 6 до 10 лет; 5) от 11 и выше. Определить моду и медиану производственного стажа рабочих.

2. Рассчитать абсолютные и относительные показатели вариации заработной платы рабочих по цехам. Сделать выводы

3. Определить количественную взаимосвязь между признаками:

3.1. С помощью графического метода определить форму связи между производственным стажем и заработной платой рабочих цеха № 2 с № 41 по № 60 включительно (п=20).

3.2. Вычислить параметры уравнения регрессии, характеризующего зависимость между производственным стажем и заработной платой рабочих. Построить на графике теоретическую и эмпирическую линии регрессии. Объяснить смысл полученных параметров уравнения.

3.3. Определить степень тесноты между рассматриваемыми признаками.

3. Методические указания по выполнению контрольных работ

Контрольная работа № 1

1.1. Ряды распределения

Простейшим примером группировки является ряд распределения, т. е. **статистическим рядом распределения** называют упорядоченное распределение единиц совокупности на группы по изучаемому признаку.

В ряду распределения различают следующие элементы: варианты и частоты, или частости. **Вариантами** (х) называют отдельные значения группировочного признака, которые он принимает в ряду распределения. Числа, которые показывают сколько раз (как часто) встречается в совокупности то или иное значение признака, или, что тоже самое, сколько единиц в совокупности обладает тем или иным значением признака, называют **частотами** f_i. **Частость** f^1_i - это относительная величина, определяющая долю частот отдельных вариант в общей сумме частот. Сумма всех частостей равна единице. Частости могут выражаться и процентах, тогда сумма всех частостей равна 100%. Термин «частость» применяется также и к показателю доли (удельного веса) единиц, обладающим определенным процентом признака, полученного по данным выборки.

Ряды распределения могут быть образованы как по атрибутивным признакам, так и по количественным. В соответствии с этим они делятся на **атрибутивные и вариационные ряды распределения.**

Вариационные ряды могут быть **дискретными** и **интервальными.**

Дискретный ряд распределения – это ряд, в котором варианты выражены одним конечным числом. **Интервальный** ряд – это ряд, в котором значение признака заданы в виде интервала. Причем, интервалом называется разность между максимальным и минимальным значением признака в каждой группе.

При построении интервальных рядов необходимо определить количество групп, величину интервала и какие взять интервалы (равные, неравные, открытые, закрытые). Эти вопросы решаются на основе экономического анализа сущности изучаемых явлений, поставленной цели и характера изменения признака.

Величина интервала определяется следующим образом:

1) когда число групп (k) оговаривается исследованием, то

$$i=\frac{x_{\max}-x_{\min}}{k},$$

где: i - величина интервала;

$x_{\max}$ - максимальное значение признака;

$x_{\min}$ - минимальное значение признака.

2) когда число групп не оговаривается исследованием, то величина интервала определяется по формуле, предложенной американским ученым Г.А. Стерджессом:

$$i-\frac{x_{\max}-x_{\min}}{1+3{,}322\lg n}$$

где: n - число наблюдений признака (объем совокупности).

В основу группировки могут быть положены атрибутивные (качественные) признаки. В этом случае число групп чаще всего определяется числом значения признака.

Группировки по атрибутивным признакам часто называют **классификациями,** т. е. это такие группировки, в основании которых положены существенные признаки и которые имеют устойчивую номенклатуру групп и подгрупп (классификация отраслей, видов производства, товаров, профессий и др.)

Ряды распределения дают возможность судить о закономерности распределения и о границах варьирования совокупности. Различные обобщающие показатели (средние, мода, медиана, дисперсия и т. д.) исчисляются на основе ряда распределения.

1.2. Средние величины

Важным средством анализа являются средние величины, в обобщенной форме характеризующие типичный уровень того или иного признака изучаемой совокупности.

Наиболее часто средний уровень значений признака исчисляется по формулам **средней арифметической.**

Средняя величина исчисляется как средняя арифметическая в тех случаях, когда имеются данные об отдельных значениях признака и о числе единиц совокупности, обладающих этим значением.

В том случае, когда совокупность небольшая и значения признака повторяются, то следует использовать **среднюю арифметическую простую,** которая определяется по формуле:

$$\bar{x} = \frac{\sum x}{n},$$

где n – число значений признака.

В том случае, когда построен ряд распределения и значения признака повторяются, то следует использовать **среднюю арифметическую взвешенную**:

$$\bar{x} = \frac{\sum xf}{f} \text{ (для дискретного ряда).}$$

Если построен интервальный ряд распределения, то следует его преобразовать в дискретный, рассчитав центральное значение признака в каждой группе, т.е.

$$\bar{x} = \frac{\sum x'f}{f}$$

Представим расчет средней арифметической в рядах распределения.

Пример №1. Имеются следующие данные, на основе которых следует определить средний стаж работы рабочих:

Группы рабочих по стажу, лет	Число рабочих f	Середина интервала x^1	$x^1 f$
3-5	10	4	40

5-7	30	6	180
7-9	40	8	320
9-11	15	10	150
11-13	5	12	60
Итого	100		750

В данном ряду варианты представлены не одним числом, а в виде интервала. Таким образом, каждая группа ряда распределения имеет нижнее и верхнее значения вариант, иначе закрытые интервалы. Расчет проводится по средней арифметической взвешенной:

$$\bar{x} = \frac{\sum x^1 f}{\sum f},$$

где: x^1 - центральное значение признака в каждой группе (середина интервала).

$$\bar{x} = \frac{\sum x^1 f}{\sum f} = \frac{750}{100} = 7{,}5 \text{ лет}$$

Преобразуем рассмотренный выше ряд распределения в ряд с открытыми интервалами.

Группы рабочих по стажу, лет	Число рабочих
До 5	10
5-7	30
7-9	40
9-11	15
11 и более	5
Итого	100

В таких рядах условно величина интервала в первой группе принимается равной величине интервала в последующей группе, а величина интервала в последней группе – величине интервала в предыдущей группе. Дальнейший расчет аналогичен изложенному выше.

Достоинством средней как обобщающего показателя является то, что она одной величиной характеризует целую совокупность различных величин. Но для всесторонней характеристики совокупности, так и для решения некоторых практических задач, нужны и такие обобщающие показатели, которые характеризуют особенности распределения единиц совокупности по величине изучаемого признака. К таким показателям относятся **мода и медиана**, которые называются **структурными средними**.

Модой в статистике называется наиболее часто встречающееся значение признака в вариационном ряду, т. е. варианта, у которой частота (вес) наибольшая.

В дискретном вариационном ряду модой будет варианта, имеющая наибольшую частоту.

В интервальном вариационном ряду, с равными интервалами в котором, как известно, не приводятся все значения признака, а указываются их группы в виде интервалов, мода исчисляется по следующей формуле:

$$M_0 = x_m + i\,\frac{f_{m_0} - f_{m_{0-1}}}{(f_{m_0} - f_{m_{0+1}}) + (f_{m_0} - f_{m_{0-1}})},$$

где M_0 - мода;

x_m x_m-нижняя граница модального интервала;

f_m - частота модального интервала

$f_{m_{0-1}}$ - частота интервала, предшествующая модальному;

$f_{m_{0+1}}$ - частота интервала, следующего за модальным;

i - величина модального интервала.

Мода имеет важное значение для решения некоторых задач, например, какой размер обуви чаще всего встречается у мужчин и женщин, какое время дня является «пиковым» для работы предприятий общественного питания, городского транспорта и др.

Медианой в статистике называется варианта, расположенная в середине вариационного ряда.

Если ряд распределения дискретный и имеет нечетное число членов, то медианой будет варианта, находящаяся в середине упорядоченного ряда (упорядоченный ряд – это ряд, в котором значения признака расположены в порядке возрастания или убывания). Например, стаж пяти рабочих составил 2, 4, 7, 8, и 10 лет. В таком ряду медиана равна 7 годам.

Если упорядоченный ряд состоит из четного числа членов, то медиана будет определяться как средняя арифметическая из двух вариант, расположенных в середине ряда. Например, пусть будет не пять, а шесть человек со стажем 2, 4, 6, 7, 8, и 10 лет. В этом случае медиана будет определяться как $\frac{6+7}{2}$, т. е. 6, 5 лет.

Если же в дискретном ряду значения признаков повторяются, то для расчета медианы используется сумма накопленных частот.

Пример №2. Имеется следующий дискретный ряд распределения рабочих по стажу работы, на основе которого следует определить медиану стажа работы рабочих:

Стаж работы, лет	Число рабочих	Сумма накопленных частот
10	2	2
12	6	8(6+2)
15	16	24(16+8)
17	12	
20	4	
Итого	40	

Для определения медианы надо подсчитать сумму накопленных частот ряда. Наращивание итога продолжается до получения суммы частот немногим больше половины В нашем примере сумма равна 40, ее половина – 20. Накопленная сумма частот ряда получилась равной 24.

Варианта, соответствующая этой сумме, т. е. 15 лет, и есть медиана данного ряда.

Смысл полученного результата следующий: одна половина рабочих имеет стаж меньше 15 лет, а другая – больше 15 лет.

Если же сумма накопленных частот против одной из вариант будет равна точно половине суммы частот, то медиана определяется как средняя арифметическая из этой варианты и последующей.

В интервальном ряду распределения с равными интервалами медиана определяется по следующей формуле:

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

где: Me - медиана;

x_{Me} - нижняя граница медианного интервала;

i - величина медианного интервала;

S_{Me-1} - сумма накопленных частот, предшествующих медианному интервалу;

f_{Me} - частота медианного интервала.

Хотя мода и медиана не получили в статистике такого широкого применения, как средние величины, их использование в некоторых экономико-статистических и других расчетах может иметь важное значение. Мода позволяет выявить наиболее распространенный размер или уровень изучаемого явления, а медиана характеризует деление совокупности явления на две половины, одна из которых ниже, а другая выше определенного уровня.

Рассмотрим расчет моды и медианы.

Пример №3. Распределение предприятий по численности работающих характеризуется следующими данными:

Группировка предприятий по числу работающих	Число предприятий	Сумма накопленных частот
100-200	1	1
200-300	3	4(3+1)
300-400	7	11(7+4)
400-500	30	41(30+11)
500-600	19	
600-700	15	
700-800	5	
Итого	80	

В этой задаче наибольшее число предприятий (30) имеет численность работающих от400 до 500 чел. Следовательно, этот интервал является модальным интервалом ряда распределения. Введем обозначения:

$$x_{Mo} = 400;\ i_{Mo} = 100;\ f_{Mo} = 30;\ f_{Mo-1} = 7;\ f_{Mo+1} = 19.$$

$$Mo = 400 + 100\frac{30-7}{(30-7)+(30-19)} = 400 + 100\frac{23}{23+11} = 468 \text{ чел.}$$

Таким образом, среди 80 предприятий наиболее часто встречаются предприятия с численностью работников равной 468 чел.

Чтобы определить медиану необходимо произвести накопление частот до величины чуть больше половины суммы всех частот. Против варианты 400-500 сумма накопленных частот равна 41, отсюда следует, что *Ме* будет находиться в этом интервале. Ее величина будет равна:

$$Me = 400 + 100\,\frac{40-11}{30} = 497 \text{ чел.}$$

Следовательно, половина предприятий имеет численность работников меньше 497 чел, а половина – более 497 чел.

1.3.Показатели вариации

Средние величины дают обобщающую характеристику уровня количественного признака по совокупности в целом. В одних случаях отдельные значения признака могут очень незначительно отличаться от средней арифметической, в других случаях, наоборот, отдельные значения далеко отстоят от средней.

Изменение (колеблемость) величины количественного признака от одной единицы однородной совокупности до другой принято называть вариацией. Размеры вариации позволяют судить, насколько однородна изучаемая группа и, следовательно, насколько характерна средняя по группе. Изучение отклонений от средних имеет большое практическое и теоретическое значение, поскольку в отклонениях проявляется развитие явления; небольшие количественные изменения, постоянно нарастая, могут в дальнейшем привести к существенным, качественным сдвигам.

Для характеристики размера вариации используются специальные показатели колеблемости (вариации): размах вариации, среднее линейное отклонение, среднее квадратическое отклонение, коэффициент вариации.

Размах вариации равен разности между наибольшим и наименьшим значениями варьирующего признака в данном ряду распределения. Это наиболее простой, но зато и наименее точный показатель вариации. При его расчете не учитываются колеблемость тех значений признака, которые заключены между его крайними значениями, и частоты разных значений признака. Наименьшее и наибольшее значение признака могут оказаться случайными, нехарактерными для совокупности и существенно отличными от других его значений.

Эти недостатки устраняются, если применить другие показатели вариации.

Среднее линейное отклонение можно рассчитать по той же формуле, что и среднюю арифметическую, но в качестве варианты выступают абсолютные отклонения (модули) значений признака от его среднего значения.

Соответственно формула расчета среднего линейного отклонения ($\overline{d}$) имеет вид:

$$\overline{d} = \frac{\sum \left| x - \overline{x} \right| \times f}{\sum f}$$

Если исчислить среднюю не из абсолютных отклонений значений признака от средней, а из квадратов этих отклонений, то получим показатель дисперсии (σ^2), квадратный корень из которого называют средним квадратическим отклонением (σ):

$$\sigma^2 = \frac{\sum (x - \overline{x})^2 \times f}{\sum f}$$ - дисперсия;

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2 \times f}{\sum f}}$$ - среднее квадратическое отклонение.

Среднее линейное отклонение также как и среднее квадратическое отклонение показывают, насколько в среднем отличаются индивидуальные признаки от среднего его значения.

Выше названные показатели являются **абсолютными показателями вариации**. Эти величины – именованные, они зависят от масштаба измерения признака, поэтому не всегда пригодны для сравнения.

Абсолютными показателями вариации нельзя непосредственно пользоваться в двух случаях:

- для сравнения степени вариации двух различных признаков в одной и той же группе;

• для сравнения вариации по одному и тому же признаку, но в двух различных группах с разным уровнем средних, т.е. значительно чаще отличающихся по объему совокупности.

Поэтому чаще используется относительный показатель вариации, который носит название коэффициента вариации (V) и рассчитывается по формуле:

$$V = \frac{\sigma}{\overline{x}} \times 100\%$$

Коэффициент вариации показывает, на сколько процентов в среднем индивидуальные значения признака отличаются от его среднего значения. В известной степени коэффициент вариации является критерием надежности средней: если он велик (более 33%), то это свидетельствует о большой колеблемости в величине признака у отдельных единиц данной группы, а, следовательно, средняя недостаточно надежна.

Рассмотрим расчет этих показателей на основе данных, представленных в таблице.

Пример № 4.

Группировка рабочих по стажу, лет	Число рабочих	x'	$x' \cdot f$	$x - \overline{x}$	$\lvert x - \overline{x} \rvert \cdot f$	$(x - \overline{x})^2 \cdot f$
До 2	22	1	22	-3	66	198
2-4	26	3	78	-1	26	26
4-6	32	5	160	1	32	32
6-8	20	7	140	3	60	180
Итого	100		400		184	436

$$\overline{x} = \frac{\sum x' \cdot f}{\sum f} = \frac{400}{100} = 4 \text{ года}$$

$$\overline{d} = \frac{\sum \lvert x - \overline{x} \rvert \times f}{\sum f} = \frac{\sum \lvert \overline{d} \rvert \cdot f}{\sum f} = \frac{184}{100} \approx 1{,}8 \quad \text{года}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \overline{x})^2 \cdot f}{\sum f}} = \sqrt{\frac{436}{100}} = 2{,}1 \text{ года}$$

$$v = \frac{\sigma}{\overline{x}} \cdot 100\% = \frac{2{,}1}{4} \cdot 100\% = 52{,}5\%$$

На основе коэффициента вариации можно сделать вывод, что средняя в данной совокупности не надежна, т.е. совокупность неоднородна.

Если совокупность разбита на группы по изучаемому признаку, то для такой совокупности могут быть исчислены следующие виды дисперсий: общая, групповые (частные), средняя из групповых (частных) и межгрупповая.

Общая дисперсия (σ^2) равна среднему квадрату отклонений отдельных значений признака (x) от общей средней ($\bar{x}$). Она может быть исчислена как простая средняя или взвешенная соответственно по формулам:

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n};$$

$$\sigma^2 = \frac{\sum (x - \bar{x})^2 \times f}{\sum f}$$

Общая дисперсия отражает вариацию признака за счет всех условий и причин, действующих в совокупности.

Групповая дисперсия (σ_i^2) равна среднему квадрату отклонений отдельных значений признака внутри группы от средней арифметической этой группы ($\bar{x}_i$). Она может быть исчислена как простая средняя или взвешенная соответственно по формулам:

$$\sigma^2 = \frac{\sum (x - \bar{x}_i)^2}{n};$$

$$\sigma^2 = \frac{\sum (x - \bar{x}_i)^2 \times f}{\sum f}$$

Эта дисперсия отражает вариацию признака за счет условий и причин, действующих внутри группы.

Средняя из групповых дисперсий ($\bar{\sigma}_i^2$) – это средняя арифметическая из дисперсий групповых:

$$\bar{\sigma}_i^2 = \frac{\sum \sigma_i^2}{n}; \quad \bar{\sigma}_i^2 = \frac{\sum \sigma_i^2 \cdot f}{\sum f}$$

Межгрупповая дисперсия (∂^2) равна среднему квадрату отклонений групповых средних от общей средней:

$$\partial^2 = \frac{\sum (x_i - \overline{x})^2}{n}$$

$$\partial^2 = \frac{\sum (x_i - \overline{x})^2 \cdot f}{\sum f}$$

Межгрупповая дисперсия характеризует вариацию результативного признака за счет признака группировочного.

Между указанными видами дисперсий существует определенное соотношение: **общая дисперсия равна сумме средней из групповых дисперсий и межгрупповой дисперсии**:

$$\sigma^2 = \overline{\sigma_i^2} + \partial^2$$

Это соотношение называют **правилом сложения дисперсий**.

Правило сложения дисперсий используется в статистике для определения степени тесноты связи между изучаемыми признаками.

С этой целью рассчитывается **коэффициент детерминации** (η^2), который представляет собой отношение межгрупповой дисперсии к общей, и показывает, какую часть общей вариации изучаемого признака составляет вариация межгрупповая, т.е. обусловленная группировочным признаком.

Корень квадратный из коэффициента детерминации называется **эмпирическим корреляционным отношением:**

$$\eta = \sqrt{\frac{\partial^2}{\sigma^2}}$$

Оно характеризует степень тесноты связи между взаимосвязанными признаками. По его абсолютной величине судят о тесноте связи или степени зависимости признака результативного от одного признака факторного (группировочного) или нескольких.

Эмпирическое корреляционное отношение может изменяться от 0 до 1. Чем ближе его величина к единице, тем связь теснее. Знак указывает на характер, направление связи.

Для совокупности с большим объемом совокупности, а также если значения признака выражены большими числами, процесс вычисления $\bar{x}$, σ, σ^2 значительно усложняется и становится нередко громоздким и трудоемким. Поэтому в статистике разработаны приемы упрощенного вычисления этих показателей. Одним из наиболее эффективных способов упрощения является способ условных моментов, который основан на использовании свойств средней арифметической и дисперсии.

Условным моментом называется средняя арифметическая из отклонений отдельных значений признака от некоторой постоянной величины A, называемой условным началом ($A \neq 0$, $A \neq \bar{x}$). Обычно за условное начало A принимают варианту наиболее часто встречающуюся в совокупности, т.е. моду.

В зависимости от степени, в которой берутся отклонения, моменты бывают разных порядков. Если средняя рассчитывается из отклонений первой степени, получается условный момент первого порядка (M_1), из отклонений второй степени – условный момент второго порядка (M_2) и т.д., т.е.

$$M_1 = \frac{\sum \left(\frac{x - A}{i} \right) \cdot f}{\sum f}; \quad M_2 = \frac{\sum \left(\frac{x - A}{i} \right)^2 \cdot f}{\sum f}$$

Используя способ условных моментов, можно рассчитать $\bar{x}$, σ, σ^2, по формулам:

$$\bar{x} = A + i + M_1;$$

$$\delta^2 = i^2 \cdot (M_2 - M_1^2);$$

$$\delta = i \sqrt{M_2 - M_1^2}$$

Пример № 5: на основе ряда распределения, характеризующего время горения электроламп, определить среднее время горения электроламп, среднее квадратическое отклонение и дисперсию времени горения электролампы, пользуясь способом условных моментов.

Группы электроламп по времени горения (час)	Число Ламп	$x^!$	$x^! - A$	$\frac{x^! - A}{i}$	$\left(\frac{x^! - A}{i}\right) f$	$\left(\frac{x^! - A}{i}\right)^2 \cdot f$
800-10000	20	900	-400	-2	-40	80
1000-1200	80	1100	-200	-1	-80	80
1200-1400	160	1300	0	0	0	0
1400-1600	90	1500	200	1	90	90
1600-1800	40	1700	400	2	80	160
1800-2000	10	1900	600	3	30	90
Итого	400				80	500

Данный ряд распределения является интервальным.

Для расчета среднего времени горения электролампы используется формула средней арифметической взвешенной:

$$\bar{x} = \frac{\sum x^! \cdot f}{\sum f}$$

где $x^!$ - центральное значение признака в каждой группе, или, пользуясь способом условных моментов:

$$\bar{x} = A + i \cdot M_1 = 1300 + 200 \cdot \frac{80}{400} = 1340 \text{ ч.},$$

где A - условное начало, т.е. A = 1300ч (варианта, которая наиболее часто повторяется;

M_1 - момент первого порядка.

Дисперсия времени горения электролампы равна:

$$\sigma^2 = i^2 \left(M_2 - M_1^2\right);$$

где M_2 - момент второго порядка.

$$\sigma^2 = 200^2\left[\frac{500}{400} - \left(\frac{80}{400}\right)^2\right] = 200^2(1,25 - 0,04) = 48400 \text{ ч.}$$

Среднее квадратическое отклонение:

$$\sigma = \sqrt{\sigma^2} = i \cdot \sqrt{M_2 - M_1^2} = \sqrt{48400} = 220 \text{ ч.}$$

Контрольная работа № 2

При статистическом изучении корреляционной связи между двумя признаками исходными данными являются данные об индивидуальных значениях признаках этих признаков в изучаемой совокупности единиц.

Корреляционные связи бывают линейные и криволинейные. Под линейной корреляционной связью понимают такую связь, при которой с возрастанием одного признака происходит непрерывное возрастание /или убывание/ другого признака в среднем на постоянную величину. Эта связь описывается уравнением прямой. При криволинейной связи между признаками имеется не постоянное, а меняющееся соотношение /результативный признак то увеличивается, то уменьшается с различной степенью интенсивности. Эта связь описывается уравнением какой-либо кривой.

2.1. Определение формы связи между признаками

Применяются несколько способов выявления наличия связи между показателями и ее формы:

1. Метод параллельных рядов /параллельного сопоставления./
2. Графический метод.
3. Способ группировки и выведения средних по группам.

Сущность **способа параллельных рядов или параллельного сопоставления** заключается в том, что факториальный признак (x) располагают в порядке возрастания и против каждого его значения записывают соответствующее значение результативного признака ($\bar{y}$). Если с увеличением одного признака другой возрастает (убывает), то между ними имеется связь. Этот анализ может предсказать и форму связи – является ли эта связь линейной или более сложной.

Сущность **графического метода** заключается в построении **поля корреляции**, представляющий собой точечный график, для построения которого в прямоугольной системе координат по оси абсцисс откладывают значения факториального признака, а по оси ординат – результативного признака. Получают в поле графика точки, соответствующие этим значениям. По тому, как располагаются эти точки, судят о наличии и форме связи.

Эти два метода наиболее простые, но громоздкие. Их целесообразно использовать в случае, если объем совокупности небольшой.

Более отчетливо корреляционная связь проявляется при использовании **метода группировок и расчета средних по группам**. Сущность этого метода заключается в том, что совокупность разбивают на группы по факториальному признаку и для каждой группы рассчитывают средние значения результативного признака. Благодаря исчислению этих средних (групповых средних) влияние прочих случайных причин, взаимопогашается и проявляется воздействие именно факториального признака. Средние значения результативного признака ($\bar{y}$) наносят на график, соединив точки, которые им соответствуют, получают эмпирическую линию связи или линию регрессии.

Когда изменение величины факториального признака в определенном направлении вызывает изменение величины результативного признака в том же направлении, то такая корреляционная связь считается прямой (положительной связью). Если же увеличение одного признака обуславливает уменьшение величины другого, находящегося с ним в корреляционной связи, или, наоборот, уменьшение величины одного признака вызывает увеличение другого, то такая корреляционная связь считается обратной (отрицательной). Этот метод является основным для выявления наличия и формы корреляционной связи.

2.2. Определение параметров уравнения связи.

Общий вид корреляционного уравнения прямой линии регрессии, т. е. уравнения прямолинейной корреляционной связи, выглядит следующим образом

$$y = a + bx,$$

где x и y - индивидуальные значения соответственно факториального и результативного признаков;

a и b - параметры уравнения прямолинейной и корреляционной связи.

Найти теоретическое уравнение связи – значит в данном случае определить параметры прямой.

Эти численные значения / параметры уравнения/ определяются на основе имеющихся данных наблюдения способом наименьших квадратов. Его сущность заключается в следующем. Теоретическая линия регрессии должна изображать изменение средних /рассчитанных как среднее арифметическое/ величин результативного признака «y» по мере изменения величин факториального признака x при условии полного взаимопогашения всех прочих, случайных по отношению к фактору x, причин. Поэтому теоретическая линия регрессии должна обладать основными свойствами средней арифметической, т. е. должна быть проведена так, чтобы сумма отклонений точек эмпирической линии регрессии от соответствующих точек теоретической линии регрессии = 0, а сумма квадратов этих отклонений была бы минимальной величиной.

Если обозначить ординаты фактических точек эмпирической линии регрессии, т. е. индивидуальные значения результативного признака, через «y_i», а ординаты теоретической линии регрессии – через «$\overline{y_x}$», следовательно, эти свойства можно представить как:

$$\sum_i^n (y_i - \overline{y_x})^2 = \min \qquad \begin{cases} \sum (y_i - \overline{y_x}) = 0 \\ \sum (y_i - \overline{y_x})^2 = \min \end{cases}$$

Это условие и лежит в основе способа наименьших квадратов.

Т.к. $\overline{y_x} = a + bx, mo \sum (y_i - a - bx)^2 = \min$.

Рассчитывая первую производную по *«a»* и первую производную по *«b»* от этой функции и приравнивая каждую из производных 0, получим возможность определить те значения *«a»* и *«b»*, при которых $\sum (y - \overline{y})^2 = \min,$, т. е. необходимо решить следующую систему нормальных уравнений

$$\begin{cases} \sum y = an + b\sum x \\ \sum xy = a\sum x + b\sum x^2 \end{cases}$$

Параметр «*b»* играет решающую роль при определении характера связи. Он показывает, насколько возрастет *«y»* при каждом возрастании *«x»* на единицу. Если зависимость положительная прямолинейная, то и параметр *«b»* является положительной величиной, если зависимость обратная, то этот параметр является отрицательной величиной.

Определив параметры a и b и подставив их значения в уравнения связи, найдем количественную характеристику связи между результативным и факториальным признаками. Подставляя в это уравнение индивидуальные значения факториального признака, определяют средние величины результативного признака для соответствующих значений факториального признака. По полученным величинам средних значений результативного признака составляют теоретическую линию регрессии, характеризующую форму корреляционной связи между изучаемыми признаками.

2.3. Определение степени тесноты и существенности связи между признаками

Степень тесноты связи можно измерить следующими показателями:

1) в случае линейной формы связи применяется линейный коэффициент корреляции τ или корреляционное отношение η в том случае, если связь прямая;

2) в случае криволинейной связи рассчитывают лишь корреляционное отношение.

Линейный коэффициент корреляции τ рассчитывается по формуле:

$$\tau = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

Коэффициент корреляции показывает не только тесноту, но и направление связи. Его значение изменяется от -1 до +1. Если коэффициент корреляции имеет знак плюс, то связь прямая. Близость к единице в том и другом случае характеризует близость к функциональной зависимости.

В случае линейной формы связи факт совпадения или несовпадения η и τ используется для оценки формы связи. Установлено, что если разность между η и τ не превышает 0,1, то гипотезу о прямолинейной форме связи можно считать подтвержденной. Если разность между η и τ больше 0,1, то на различие следует обратить внимание, т. е. связь не является абсолютно прямолинейной.

Оценка существенности коэффициента корреляции определяется на основании критерия его надежности t, который рассчитывается по формуле:

$$t = \frac{|\tau|\sqrt{n-1}}{1-\tau^2},$$

где n - объем совокупности.

В математической статистике доказано, что если критерий надежности численно меньше, чем величина 2,56, то связь между коррелируемыми признаками признается несущественной. В этом случае считается, что признак-фактор не оказывает существенного влияния на результативный признак. Если t > 2,56, то связь между признаками считается существенной, а, следовательно, факториальный признак оказывает существенное влияние на результативный признак.

Литература

1. Теория статистики/ под ред. проф. Л.Г.Громыко. - М.:Инфра-М., 2011.
2. Громыко Г.Л. Теория статистики: практикум. - М.: Инфра-М., 2011.

№	Программное обеспечение и интернет-ресурсы
3.	Электронные ресурсы библиотеки Университета – электронные версии пособий, методических разработок, указаний и рекомендаций по всем видам учебной работы.