Национальный исследовательский университет «Высшая школа экономики»

Факультет: Московский институт электроники и математики

Образовательная программа: Прикладная математика

Отчет по Домашнему заданию № 1 «Пространственные данные»

по майору «Прикладной статистический анализ»

Работу выполнил

студент 2 курса:

Цыплаков Александр

Александрович

Преподаватель:

Кабаева Елена

Владимировна

Содержание

Введение

Для анализа был выбран массив исходных данных – «Количество персонала в крупных компаниях России на 2018г.» (Приложение 1). Тематически эти данные обоснованы сферой моих интересов – статистическим исследованием крупных компаний. Существует достаточно мало работ, в которых проводится анализ статистических показателей, связанных с численностью персонала. Работая с выбранным массивом, можно наблюдать за распределением персонала в крупных компаниях.

Источником исходных данных является статья, выпущенная частной газетой «Управление производством».

Данные были подобраны с учетом заданных характеристик. Таким образом, данные в исходном ряду являются по своему типу пространственными (не временными); по своему характеру первичными (т.е. отражают непосредственно значения наблюдений), одномерными и моментными(не интервальными), сопоставимыми. Они представляют собой абсолютные значения (т.е. не являются расчетными из других показателей), разрывы в значениях отсутствуют, данные одного порядка.

Первоначально исходные данные представляли собой 59 различных значений по данным персонала в крупнейших компаниях. Позже, на шаге расчетов интервалов и построения интервального вариационного ряда, исходные данные были скорректированы – были исключены выбросы (самые маленькие и самые большие значения ранжированного ряда) для корректировки интервалов с целью избежать наличия таких, частота которых является нулевой (подробнее в главе «Построение интервального вариационного ряда»). После данного изменения величина массива данных составила 44 измерений, с этим количеством были проведены все основные расчеты. Для расчетов использована программа Numbers для Mac.

Построение интервального ряда

В первую очередь при анализе имеющейся базы данных необходимо построить интервальный вариационный ряд признака распределения. Изначально ряд был построен по необработанным данными, в связи с чем возникла проблема, некоторые интервалы содержали нулевое количество значений, и именно после этих вычислений было принято решение отредактировать данные и исключить данные с максимальными и минимальными значениями, так как их можно было считать выбросами. Для максимальных, так как они намного превосходили значения, идущие перед ними, для минимальных, так как они были значительно меньше других значений и исчислялись не в тысячах персонала, а в единицах. Соотвественно , все последующие вычисления были проведены с отредактированным рядом чисел, включающим в себя 44 элемента.

Для построения интервального вариационного ряда была построена расчетная таблица. С помощью функции «Сортировка» было проведено ранжирование исходных данных в порядке возрастания. Ряд ранжированных данных представлен в столбце В.

Далее мной были определены минимальное и максимальное значения ряда данных. Для этого были применены функции МИН и МАКС соответственно, после чего я сравнил эти два значения с крайними значениями ранжированного ряда и убедился в правильности результата, таким образом: Xmin = 15 000 (человек), Xmax = 113 582 (человек). Исходя из полученных данных можно определить размах вариационного ряда, равный разности между максимальным и минимальным значением, получился 98 582(человек).

Затем по заданию необходимо рассчитать количество интервалов. Для нахождения оптимального числа интервалов были использованы два способа: формула Брукса и Каррузера (n=5*log(N)) и правило Стерджесса (n = 1 + 3,322*log(N), где n - число интервалов, N - число наблюдений. По первой из представленных формул n = 6 (после округления), по второй - 8 (после округления). После анализа всех возможных вариантов, основным был выбран n = 7, величина, находящаяся между двумя полученными значениями.

Для определения шага, иначе говоря ширины, интервала была использована формула h = R/n, где h – ширина интервала, R – размах вариационного ряда. Значение ширины интервала получилось равным 16 430 (человек). Это значение использовалось на протяжении всей оставшейся работы. Все интервалы в моем анализе не имеют нулевого количества значений.

Далее при расчёте границ каждого интервала использовались формулы a1 = Xmin - h/2; a2=b1=a1 + h; a3=b2=a2 + h, где ai - нижняя граница i-го интервала, bi - верхняя граница i-го интервала. При этом верхние границы каждого интервала были подвергнуты корректировке, а именно после получения результата в формулах, представленных выше, было сделано еще одно действие: -1, дабы исключить совпадения верхней границы и нижней границы следующего интервала, чтобы одно и то же значение не попало в два интервала. Кроме того, параллельно с этим были вычислены середины интервалов путем полуразность верхней и нижней границы каждого. Тем самым мы получили интервалы, в которых содержаться все значения нашего датасета.

Таким образом, исходный ряд, состоящий из 44 наблюдений, был преобразован в интервальный вариационный ряд с 7 равными интервалами. Шаг каждого интервала составляет 16 430 человек. Нижняя граница интервального ряда составляет 6 785 человек. Верхняя граница - 121 797.

Для каждого из полученных интервалов были вычислены частота и накопленная частота. Частота - количество вхождений данных в каждый интервал. Максимальная частота в интервале 23 215 человек - 39 645 человек и составляет - 14. Это означает, что 14 компаний из 44 на 2018 год в среднем имели персонал 31 430 человек. Минимальная частота содержится сразу в 2х интервалах: 1)88 936 человек - 105 366 человек; 2)105 367 человек - 121 797 человек; и составляет - 1. Это означает, что на 2018 год только по одной компании в среднем имели персонал 97 151 человек и 113 582 человек.

Отредактированный интервальный ряд (нижняя и верхняя границы каждого интервала), значение середины для каждого интервала, частота и накопленная частота представлены в Приложении 2.

Визуальное представление данных

Для визуального представления данных были построены гистограмма, полигон, комулята и огива по данным интервального ряда.

Гистограмма (Рис. 1) отражает в виде столбчатой диаграммы число элементов общей выборки в каждом отдельно взятом интервале Для описанных выше данных на оси абсцисс отражается каждый интервал, на оси ординат - число компаний, попавших в интервал. Так, к примеру, в интервале от 6 785 до 23 214 человек персонала располагается 12 компаний из выборки.

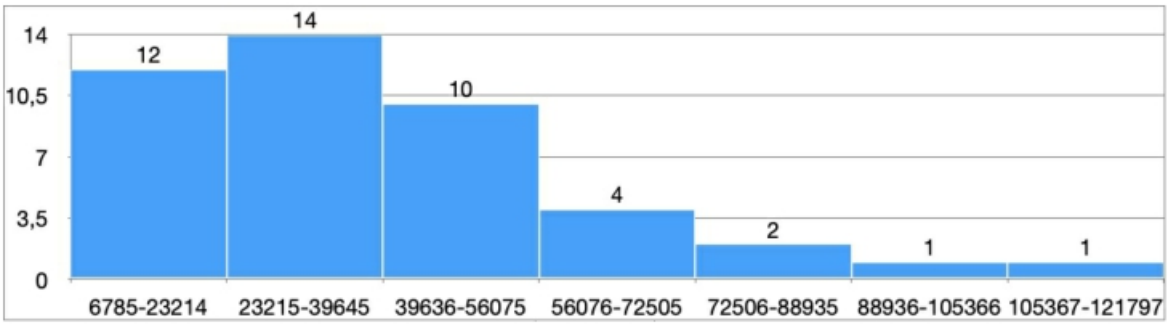

Рис. 1 Гистограмма распределения компаний по численности персонала на 2018 год.

Полигон (Рис. 2) представляет из себя графическое изображение варационного ряда в виде соединяющей точки ломанной. По оси абсцисс отмечены точки, обозначающие границы каждого интервала, по оси ординат - частоту каждого интервала. Этот вид визуализации помогает отразить, что в интервале от 22 215 до 39 645 человек персонала располагается 14 компаний.

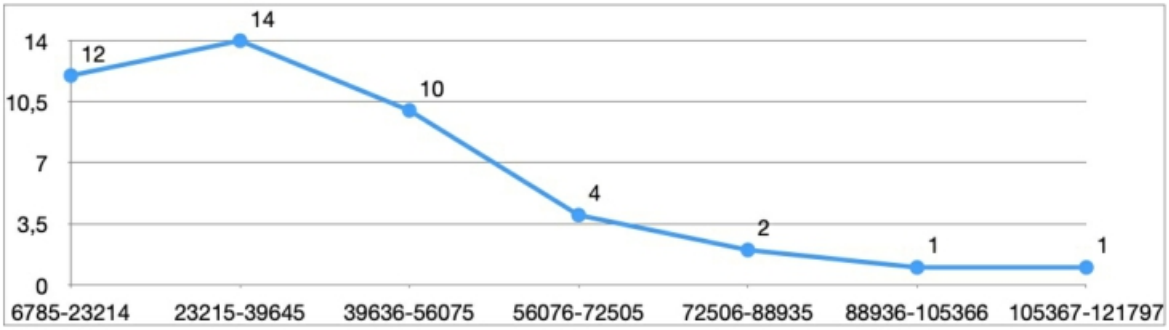

Рис. 2 Полигон распределения компаний по численности персонала на 2018 год.

Кумулята (Рис. 3) частот интервального ряда распределения графически изображает в виде столбцов накопленные частоты, относящиеся к каждому

конкретному значению. Она показывает, какое количество значений признака из первоначальной выборки не превышают заданного значения. На горизонтальной оси лежат нижние границы всех интервалов, на вертикальной - накопленные частоты. Таким образом, в описанных данных, например, 42 компании не превышают по количеству персонала 88 936 человек.

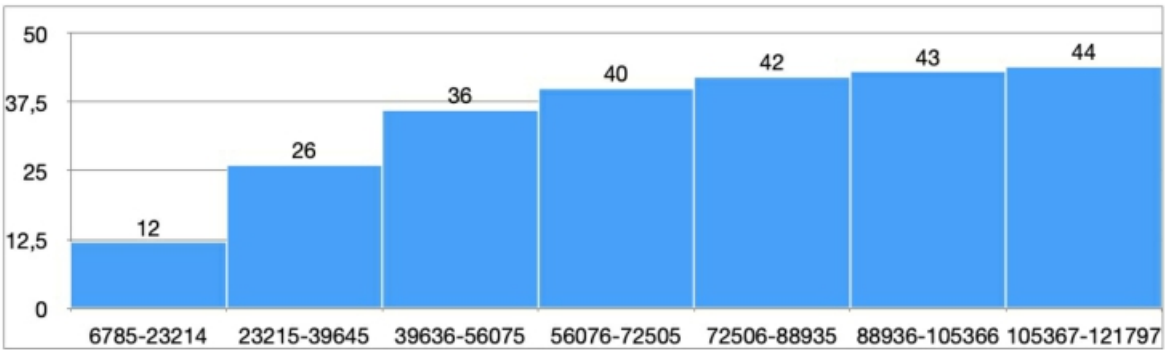

Рис. 3 Кумулята частот интервального ряда распределения компаний по численности персонала на 2018 год.

Огива (Рис. 4) строится по тем же данным, что и кумулята, разница заключается в том, что она представляет собой не столбчатую диаграмму, а ломанную, соединяющую точки, и значения осей поменяны местами (на оси абсцисс - накопленные частоты, на оси ординат - нижние границы полученных интервалов). На описанном примере прямая не является диагональю квадрата, из чего можно сделать вывод о том, что исходные данные неравномерно распределены по интервалам. К примеру, 43 из 44 компаний имеют персонал меньше 105 367 человек и отражают частоту 6 из 7 интервалов.

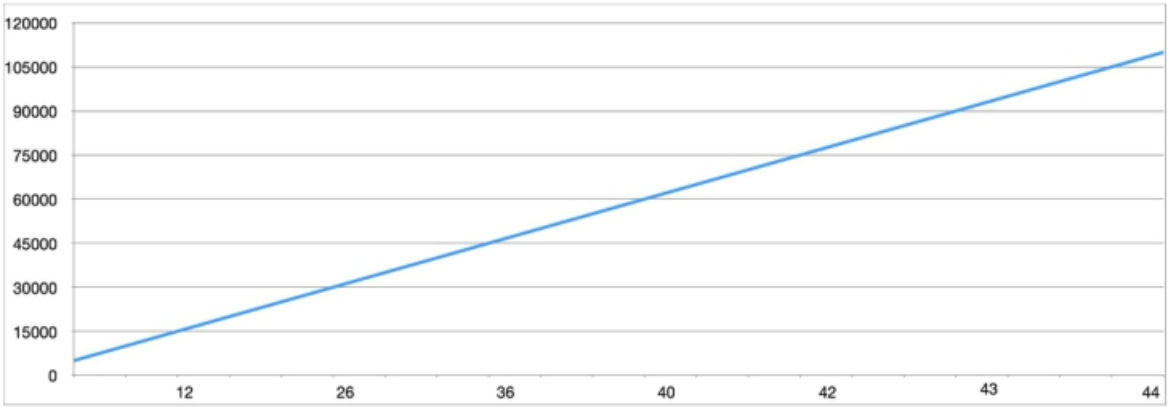

Рис. 4 Огива частот интервального ряда распределения компаний по численности персонала на 2018 год.

Характеристики центра положения

В рамках нахождения основных характеристик одномерного ряда были определены три средние величины: средняя арифметическая , средняя геометрическая и средняя гармоническая. Вычисления этих и последующих величин расписаны в расчётной таблице. Средняя арифметическая (средняя отклонений) была определена несколькими способами по дискретному вариационному ряду: одноименная формула и функция СРЗНАЧ (x = 39 864), вручную - сумма всех значений / количество значений (x = 39 864). Таким образом, оба способа расчёта дали одинаковые результаты.

Кроме того, была определена взвешенная средняя арифметическая как частное от деления суммы произведений значений и частот на число наблюдений. Ее значение равно простой средней арифметической, так как частота каждого отдельного наблюдения в исходном ряду равна единице. Этот показатель определялся по сгруппированному ряду, с использованием расчета частот - функция ЧАСТОТА.

Значение средней арифметической для интервального вариационного ряда составило - 39 272. Отличие между значениями для дискретного и интервального рядов объясняется погрешностями, возникающими при округлении значений и на этапе преобразовании дискретного ряда в интервальный.

Следующий найденный показатель - средняя геометрическая. Она была определена двумя способами для дискретного вариационного ряда: с помощью формулы и функции СРГЕОМ (x = 34 277), вручную - посредством извлечения корня 44 степени из произведения всех 44 данных (x = 34 277). Оба способа расчёта дали одинаковый результат.

Далее была найдена средняя гармоническая для дискретного вариационного ряда. Для этого были использованы следующие способы: одноименные формула и функция СРГАРМ (x = 30 196), вручную - как частное количества значений и суммы обратных значений признака для всех данных. Оба способа дали одинаковый результат. Такой же результат был получен для взвешенной средней геометрической, так как частота каждого значения равняется единице.

Значение средней гармонической для интервального вариационного ряда составило 29 283. Разница в полученных значениях объясняется погрешностями преобразования дискретного ряда в интервальный.

Следующим этапом была определена медиана, значение, которое делит элементы выборки на две равные половины. Она была определена с помощью функции МЕДИАНА (Me = 31 839) и вручную для ряда с четным числом элементов (N = 44) - путем полусуммы 22-го и 23-го значений. Способы дали одинаковые результаты. Получается, что ровно половина компаний имеют аудиторию меньше 31 839 человек на 2018 год и ровно половина компаний больше данного значения.

Также был определен медианый интервал - интервал, накопленная частота которого первой превышает половину сумм всех частот (mi > 22). Таковым является интервал, накопленная частота которого равна 26. Его нижняя граница - 23 215, а верхняя - 39 645.

Далее была определена мода - наиболее часто встречающееся значение ряда. В выбранных данных мода отсутствует (это подтверждается и расчетами вручную, и расчётами с использованием функции МОДА), так как каждое значение ряда является уникальным и имеет частоту равную единице.

Был определен модальный интервал - интервал, имеющий наибольшую частоту. Таковым интервалом является второй с частотой 14, то есть в него входят 14 из 44 компаний. Модальный ряд совпадает с медианным рядом и имеет такие же границы.

После проведения расчётов были построены графики для визуализации полученных данных. Был построен полигон распределения компаний по количеству персонала (Рис. 5), где были отмечены медиана и средняя арифметическая, мода отсутствует, поэтому отмечена не была. Условие симметрии соблюдено не было, x <> Me <> Mo, можно говорить о том, что распределение не симметрично. Me < x, следовательно асимметрия правосторонняя относительно центра распределения / левосторонняя скошенность, пик сдвинут влево

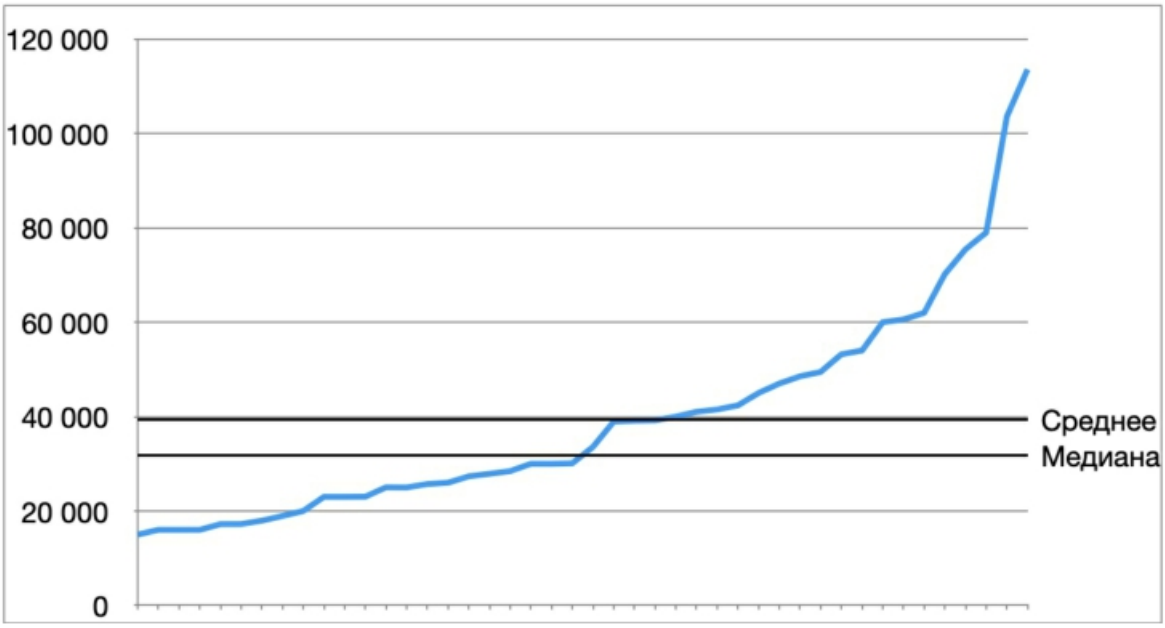

Рис. 5 Полигон распределения числа компаний, имеющих соответствующие количество персонала на 2018 год.

Кроме того, была построена гистограмма интервального ряда распределения компаний (Рис. 6), на которой были отмечены медианный и модальный интервал, они совпадают. Пик сдвинут влево, то есть наблюдается левосторонняя скошенность; правосторонняя асимметрия.

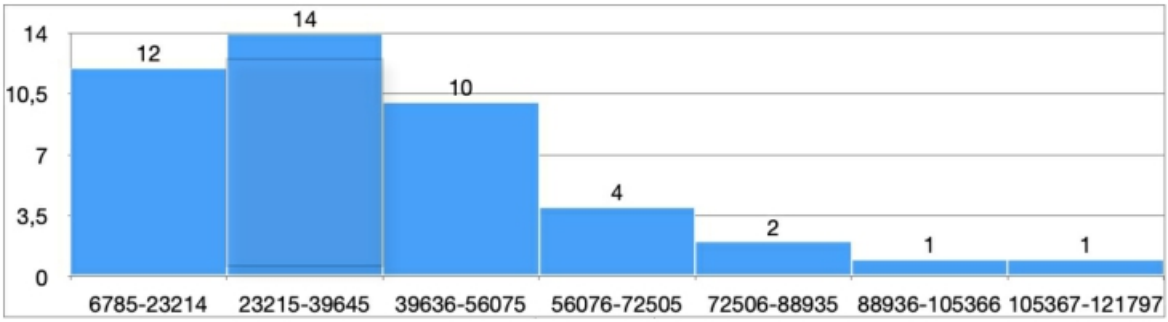

Рис. 6 Гистограмма интервального ряда распределения числа компаний, имеющих соответствующие количество персонала на 2018 год.

Коэффициенты асимметрии и эксцесса. Квартили и децили.

Следующим шагом были вычислены два коэффициент - асимметрии и эксцесса. Коэффициент асимметрии был рассчитан единственным способом - вручную по формуле, встроенной функцией СКОС воспользоваться не удалось, так как в программе Numbers она отсутствует. После вычислений получилось As = 1.

Следовательно, можно сделать вывод, что распределение не симметрично, пик сдвинут влево. Можно говорить о правосторонней асимметрии; левосторонней скошенности, то есть распределение персонала компаний склоняется к более низкому по сравнению со средним значением.

Коэффициент эксцесса также был посчитан только вручную по формуле по той же причине. Он получился равен 1. На графике распределения будет выраженный пик, так как Ек > 0. Значение далеко от нуля, значит, распределение сильно отличается от нормального.

Далее были вычислены квартире и децили для дискретного вариационного ряда, 1-ый и 3-ий квартили и 1-ый и 9-ый децили для интервального вариационного ряда. Для расчета квартилей - значений изучаемого признака, левее и правее которых находится четверть всех имеющихся наблюдений, - по исходным данным были проведены расчеты вручную.

Таблица 1 Значения квартилей для дискретного вариационного ряда вычисленные по формуле.

Квартиль	**Значение**
Q1	20 000
Q2	31 839
Q3	45 000

Также были расчитаны децили для дискретного вариационного ряда. Для этого я воспользовался формулой $D_j = j*(n-1)/10$, где при получении нецелого номера наблюдений для расчета дециля использовалось среднее между номерами наблюдениями, в которое попадает полученное число. Результаты вычислений приведены ниже (Таблица 2):

Таблица 2 Значения децилей для дискретного вариационного ряда вычисленные по формуле.

Дециль	Значение
D1	16 002
D2	17 955
D3	23 000
D4	25 700
D5	30 005
D6	38 934
D7	41 510
D8	48 527
D9	60 587

Далее по формулам вручную поэтапно были вычислены квартире и децили для интервального вариационного ряда (расчеты приведены в расчетной таблице). Итоговыми рассчитанными значениями для квартилей стали следующие: Q1 = 7 959, Q3 = -216 582. Итоговыми рассчитанными значениями для децилей стали следующие: D1 = 12 809, D3 = -44 143. Верхний квартиль и 9-ый дециль получились с отрицательными значениями, в связи с сильными отклонениями в данных. Небольшая разница обусловлена округлениями, происходящими на этапе формирования интервального вариационного ряда.

Характеристики вариации

На следующем этапе анализа были рассчитаны характеристики вариации по исходным данным и по интервальному ряду (все промежуточные результаты и ход расчетов приведены в расчетном файле). Сводная таблица со всеми показателями вариации признака по дискретному и интервальному вариационному рядам приведены в Приложении 3. Сначала была рассчитана дисперсия - мера разброса данных вокруг средней арифметической - двумя способами для исходных данных: встроенной функцией ДИСПР (S^2 = 506 266 169) и вручную по формуле (S^2 = 506 266 169). Мы видим, что значения совпали, значит вычисления проведены верно. Значение дисперсии по интервальному ряду равно 533 638 595, значение немного отличается за счет округлений, происходящих при переходе от дискретного вариационного ряда. Таким образом средний квадрат отклонений выбранных данных составляет более 500 000 000.

Далее было определено среднее линейное отклонение - показатель, дающий обобщающую характеристику размаха значений признака. Для дискретного ряда оно равняется 17 100 человек при обоих способах вычисления: по формуле и используя функцию СРОТКЛ. Для интервального ряда данный показатель составил 18 230, что можно обьяснить округлением промежуточных расчётов. Данный показатель говорит о том, что конкретные значения персонала одной компании в среднем отклоняются от среднего значения на 18 230 человек.

Среднее квадратичное отклонение - показатель рассеивания значений случайной величины относительно ее математического ожидания - для дискретного ряда при расчете вручную составило 22 500 человек при расчетах вручную, а при расчетах по формуле составил 22 760, что является незначительной погрешностью за счет округления значений на промежуточных этапах. Для интервального ряда оно равно 23 100, что можно обьяснить округлением на промежуточных этапах вычислений. Этот показатель говорит о том, что случайная величина, отражающая количество персонала компаний рассеивается примерно на 22 500 человек, относительно среднего значения.

Следующим значением был рассчитан коэффициент осцилляции. Для ряда исходных данных он составил 250%, по интервальному - 251%. Этот показатель говорит о том, что колеблемость крайних значений признака вокруг средней арифметической составляет 250%.

Также был рассчитан коэффициент линейной вариации. Его значение при расчетах для исходных данных - 43%, для интервального ряда - 46%. Небольшое различие в результатах обуславливается округлениями в промежуточных расчётах. Этот показатель описывает то, что отношение среднего линейного отклонения к средней величине в описываемых данных составляет около 44%.

Кроме того, были произведены расчеты для вычисления коэффициента вариации. По исходному ряду он составил 57%, для интервального ряда - 59%. Расхождение в 2% объясняется округлением на промежуточных этапах вычисления. Данный показатель говорит о том, что в данных по персоналу компаний доля среднего разброса случайной величины в средней величине составляет около 58%. Это относительная мера среднего разброса значений в выбранной статистической совокупности. Значение коэффициента достаточно большое, в связи с чем можно говорить о большом разбросе и маленькой выравненности исследуемых значений. Коэффициент вариации больше 33%, что говорит о неоднородности информации, значительной изменчивости вариационного ряда.

Кроме того, был рассчитан относительный квартальный показатель вариации. Его значение для исходных данных составило 39%. Для интервального ряда он составил -321% из-за отрицательного значения 9-го квартиля для интервального ряда. Значение данного показателя говорит о низкой однородности совокупности.

Заключение

В данной работе были проанализированы различные статистические показатели для численности персонала компаний Российской Федерации на 2018 год. Был построен интервальный вариационный ряд, данные были визуализированы, были произведены расчеты характеристики центра положения, коэффициенты асимметрии и эксцесса, квартире и децили, характеристики вариации.

Выводы по каждому отдельному шагу анализа были приведены после представления результатов вычислений. Если говорить о выводах в целом, по полученным результатам можно судить о том, что распределение количества персонала в компаниях неравномерное. Можно говорить о большом разбросе и маленькой выравненности исследуемых значений. Информация относительно неоднородна. Распределение умеренно ассиметричное, наблюдается левосторонняя скошенность и правосторонняя асимметрия. Пик сдвинут влево.

Можно говорить о том, что среднее количество персонала компаний составляет 39 459 человек, разброс значений большой. Исходя из этого значения, можно судить о том, насколько крупная компания в списке наших данных.

Приложения

Приложение 1

Количество персонала в крупных компаниях Российской Федерации по состоянию на 31 декабря 2018 года, человек.

№	Название компании	Количество персонала
1	Сургутнефтегаз	15 000
2	Лукойл	16 001
3	Норильский никель	16 000
4	Группа УГМК	16 002
5	Evraz	17 220
6	UC Rusal	17 232
7	Локомотивные технологии	17 955
8	Мечел	18 943
9	Татнефть	20 000
10	НЛМК	23 001
11	Северсталь	23 000
12	Металлоинвест	23 035
13	Т Плюс	25 020
14	Ташир	25 000
15	Лента	25 700
16	Группа Альфа-банк	26 000
17	Группа ГАЗ	27 344
18	VEON (Vimpelcom)	27 886
19	Трансмашхолдинг	28 448
20	Мегафон	30 005
21	ТМК	30 000

№	Название компании	Количество персонала
22	СУЭК	30 095
23	Евросибэнерго	33 583
24	Мостотрест	38 934
25	Связной	39 126
26	Синара	39 200
27	МТС	40 000
28	Сибур	41 000
29	Мираторг	41 510
30	Еврохим	42 366
31	ЧТПЗ	45 000
32	Киевская площадь	47 000
33	ОМК	48 527
34	О'Кей	49 462
35	Группа Черкизово	53 200
36	Спортмастер	54 000
37	Полюс	60 000
38	ММК	60 587
39	Илим	61 976
40	Фосагро	70 200
41	Стройгазмонтаж	75 452
42	КДВ Групп	79 000
43	Стройгазмонтаж	103 600
44	Протек	113 582

Источник: https://up-pro.ru/library/production_management/productivity/chislennost-2018/

Приложение 2

Сгруппированный интервальный вариационный ряд распределения персонала среди компаний РФ на 2018 год.

Порядковый номер интервала	Нижняя граница интервала ai	Верхняя граница интервала bi	Количество компаний (частота mi)	Количество компаний (Накопленная частота mi(H))	Середина интервала
1	6 785	23 214	12	12	15 000
2	23 215	39 645	14	26	31 430
3	39 646	56 075	10	36	47 860
4	56 076	72 505	4	40	64 291
5	72 506	88 935	2	42	80 721
6	88 936	105 366	1	43	97 151
7	105 367	121 797	1	44	113 582
Сумма			44		

Приложение 3

Показатели вариации (характеристики рассеивания) для дискретного ряда и интервального вариационного ряда.

	Для дискретного ряда		**Для интервального ряда**
	Формула	**Функция**	
Дисперсия	506 266 169	506 266 169	533 638 595
Среднее Линейное отклонение	17 100	17 100	18 230
Среднее квадратичное отклонение	22 500	22 760	23 100
Коэффициент осцилляции (%)	250	Отсутствует	251
Коэффициент линейной вариации (%)	43	Отсутствует	46
Относительный квартильный показатель вариации	39	Отсутствует	-321
Коэффициент асимметрии	1	Отсутствует	1
Коэффициент эксцесса	1	Отсутствует	1