Государственное образовательное учреждение высшего профессионального образования «Санкт-Петербургский государственный политехнический университет»

Факультет экономики и менеджмента

Кафедра «Предпринимательство и коммерция»

Курсовой проект

по дисциплине «Статистика»

на тему «Анализ эмпирического распределения»

Выполнила студентка

Щербакова Екатерина Васильевна

Санкт-Петербург 2009

CОДЕРЖАНИЕ

ВВЕДЕНИЕ

1. ТАБЛИЧНОЕ И ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ВАРИАЦИОННОГО РЯДА

2. ХАРАКТЕРИСТИКА ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ РАСПРЕДЕЛЕНИЯ

3. ОЦЕНКА ВАРИАЦИИ ИЗУЧАЕМОГО ПРИЗНАКА

4. ХАРАКТЕРИСТИКА СТРУКТУРЫ РАСПРЕДЕЛЕНИЯ

5. ХАРАКТЕРИСТИКА ФОРМЫ РАСПРЕДЕЛЕНИЯ

6. СГЛАЖИВАНИЕ ЭМПИРИЧЕСКОГО РАСПРЕДЕЛЕНИЯ. ПРОВЕРКА ГИПОТЕЗЫ О ЗАКОНЕ РАСПРЕДЕЛЕНИЯ

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

ВВЕДЕНИЕ

Статистические ряды распределения являются одним из наиболее важных элементов статистики. Они представляют собой составную часть метода статистических сводок и группировок, но, по сути, ни одно из статистических исследований невозможно произвести, не представив первоначально полученную в результате статистического наблюдения информацию в виде статистических рядов распределения.

Первичные данные обрабатываются в целях получения обобщенных характеристик изучаемого явления по роду существенных признаков для дальнейшего осуществления анализа и прогнозирования; производится сводка и группировка; статистические данные оформляются с помощью рядов распределения в таблицы, в результате чего информация представляется в наглядном рационально изложенном виде, удобном для использования и дальнейшего исследования; строятся различного рода графики для наиболее наглядного восприятия и анализ информации. На основе статистических рядов распределения вычисляются основные величины статистических исследований: индексы, коэффициенты; абсолютные, относительные, средние величины и т.д., с помощью которых можно проводить прогнозирование, как конечный итог статистических исследований.

Актуальность данной темы обусловлена тем, что статистические ряды распределения являются базисным методом для любого статистического анализа. Понимание данного метода и навыки его использования необходимы для проведения статистических исследований.

Основной целью написания курсовой работы является изучение методики статистического анализа рядов распределения. Для достижения поставленной цели были поставлены и выполнены следующие основные задачи:

1. Освещено понятие и виды статистических рядов распределения, и основные формы их представления.
2. Рассчитаны и проанализированы показатели, характеризующие центральную тенденцию, вариацию, структуру и форму ряда распределения.
3. Проведено сглаживание эмпирического распределения и проверены гипотезы о законе распределения.

В качестве данных для анализа в курсовой работе были использованы данные о распределении регионов Российской Федерации по количеству легковых автомобилей на 1000 чел. населения за 2005 г. (из статистического сборника «Регионы России» за 2007 г.).

Методической базой для написания курсовой работы является учебное пособие Н.В. Куприенко «Статистика. Методы анализа распределений. Выборочное наблюдение». Также при написании работы были использованы и учебные пособия по основам статистики следующих авторов: И.И. Елисеевой, В.М. Гусарова, М.Р. Ефимовой и др.

статистический ряд распределение сглаживание

1. ТАБЛИЧНОЕ И ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ

ВАРИАЦИОННОГО РЯДА

Статистический ряд распределения представляет собой упорядоченное распределение единиц изучаемой совокупности на группы по определенному варьирующему признаку. Он характеризует состав (структуру) изучаемого явления, позволяет судить об однородности совокупности, закономерности распределения и границах варьирования единиц совокупности[[1]](#footnote-1).

Ряды распределения подразделяют:

1. Атрибутивные – строятся по описательным признакам в порядке возрастания или убывания наблюденных значений признака. Примером атрибутивных рядов могут служить распределения населения по национальности, по профессиям, по полу; распределение предприятий по формам собственности.
2. Вариационные – строятся по количественным признакам, например, распределение рабочих по уровню квалификации, по заработной плате, распределение студентов по успеваемости.

Вариационные ряды делятся на дискретные (в которых признак принимает только целочисленные значения) и интервальные (построены на признаках, которые могут принимать любые, в том числе и дробные значения).

Вариационные ряды распределения состоят из двух элементов: вариантов и частот. Вариантом называется числовое значение количественного признака в ряду распределения. Они могут быть положительными и отрицательными, абсолютными и относительными.

Частоты – это численности отдельных вариантов или каждой группы вариационного ряда, т.е. числа, показывающие, как часто встречаются те или иные варианты в вариационном ряду. Сумма всех частот равна общему количеству элементов совокупности и называется объемом совокупности.

Частости – это частоты, выраженные в виде относительных величин (в долях единицы или процентах). Сумма частостей равна единице или 100%.

Первым шагом к упорядочиванию первичного ряда данных является его ранжирование, т.е. расположение всех его данных в возрастающем или убывающем порядке.

Способы построения дискретного и интервального вариационных рядов различны, поэтому, вначале нужно определить к какому типу относится исследуемый ряд данных. Поскольку исходными данными для анализа являются данные о числе легковых автомобилей на 1000 человек населения, то вариационный ряд будет интервальным (так как значения признака могут быть дробными).

Следовательно, в первую очередь необходимо определить количество групп и интервалы группировки.

Интервал – количественное значение, отделяющее одну единицу (группу) от другой, т.е. интервал очерчивает количественные границы групп[[2]](#footnote-2).

Ориентировочно определить оптимальное количество групп с равными интервалами можно по формуле Стерджесса:

, (1.1)

где N – численность единиц совокупности.

Формула Стерджесса пригодна при условии, что распределение единиц совокупности по данному признаку приближается к нормальному и при этом применяются равные интервалы в группах.

Численность единиц совокупности равна 80, следовательно по формуле Стерджесса количество групп вариационного ряда будет равно:



Однако поскольку формула Стерджесса дает приемлемые результаты только при анализе больших совокупностей, рассмотрим несколько различных вариантов распределения с различным количеством интервалов.

Ниже приведены таблицы вариационного ряда, построенные с использованием разного количества интервалов (рис. 1.1.).

Выбирая окончательный вариант табличного представления вариационного ряда из представленных вариантов, остановимся на первом – n=8. При n=13 наблюдает много малонаполненных групп и нулевой интервал, при n=10 также имеются малонаполненные группы.

В таблицах первая непоименованная графа (From To) содержит интервалы значений признака «Количество легковых автомобилей на 1000 чел. населения».

Второй столбец «Count» – абсолютные частоты (fi), т.е. число единиц совокупности, обладающих указанным значением признака.

Cumulative Count – накопленные абсолютные частоты, получаемые последовательным суммированием частот по группам. Сумма накопленных частот по каждой строке означает, какое количество единиц совокупности (регионов) имеет значение признака, не превышающее значения верхней границы данного интервала. Общая сумма накопленных частот соответствует объему изучаемой совокупности (80).

Percent – частости (относительные частоты, wi; выражаются в процентах), рассчитываются:

, (1.2)

где:fi – число единиц i-той группы;  – общее число единиц в совокупности; wi – доля каждой группы в общем объеме совокупности.

а)

б)

в)

Рис. 1.1. Распределение регионов России по значению показателей «Количество легковых автомобилей на 1000 человек населения» в 2005 г. с числом интервалов а) n=8, б) n=10, в) n=13

Cumulative percent – накопленные частости – это результат последовательного суммирования относительных частот по группам, итоговая сумма, очевидно, равна 100%.

Табличное представление вариационного ряда позволяет получить подробную информацию о составе и структуре изучаемой совокупности, т.е. определить какое количество единиц изучаемой совокупности обладает тем или иным значением признака и какова доля этой группы единиц в общем объеме совокупности, а также выявить закономерность изменения частот.

Из таблицы видно, что наибольшую частоту (32 или 40,0% от всего объема совокупности) имеет интервал 140,8-171,2 автомобилей на 1000 чел. населения.

Наименьшую частоту (2 или 2,5%) имеет первый интервал – 19,3-49,69 автомобиля на 1000 чел. населения.

Для более наглядного представления вариационного ряда используют статистические графики.

Статистический график представляет собой чертеж, на котором при помощи условных геометрических фигур (линий, точек или других символических знаков) изображаются статистические данные. В результате этого достигается наглядная характеристика изучаемой статистической совокупности.

Правильно построенный график делает статистическую информацию более выразительной, запоминающейся и удобно воспринимаемой[[3]](#footnote-3).

Традиционно для изображения вариационных рядов распределения в отечественной практике используются графики: гистограмма, полигон, кумулята.

На рис. 1.2 представлен полигон распределения регионов России по количеству легковых автомобилей на 1000 чел. населения за 2005 г. в абсолютных частотах при количестве интервалов n=8. Он показывает, что наибольшую частоту имеет интервал 140,8-171,17, т.е. это модальный интервал.

Рис. 1.2. Полигон распределения регионов России по количеству легковых автомобилей на 1000 чел. населения за 2005 г. в абсолютных частотах (n=8)

На рис. 1.3 приведена кумулята распределения в абсолютных частотах, а на рис. 1.4. – в относительных частотах.

Из рисунка 1.2 видно, что середина распределения приходится на интервал 140,8-171,17, следовательно, этот интервал является медианным.

Рис. 1.3. Кумулята распределения регионов России по количеству легковых автомобилей на 1000 чел. населения за 2005 г., n=8 (абсолютные частоты)

Одной из часто используемых видов графиков является гистограмма (или столбиковая диаграмма), т.е. график распределения, на котором частоты каждого интервала представлены в виде столбиков (рис. 1.5).

Рис. 1.4. Кумулята распределения регионов России по количеству легковых автомобилей на 1000 чел. населения за 2005 г., n=8 (относительные частоты)

Рис. 1.5. Гистограмма распределения количества легковых автомобилей на 1000 чел. населения по регионам России за 2005 г. (n=8)

2. ХАРАКТЕРИСТИКА ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

РАСПРЕДЕЛЕНИЯ

Статистический анализ вариационных рядов распределения предполагает расчет характеристик центра распределения, его структуры, оценку степени вариации и дифференциации изучаемого признака, изучение формы распределения.

В качестве показателей центральной тенденции распределения используются: среднее арифметическое значение, мода и медиана.

Средней арифметической величиной называется такое значение признака в расчете на единицу совокупности, при вычислении которого общий объем признака в совокупности сохраняется неизменным.

Иными словами, средняя арифметическая величина — среднее слагаемое. При ее вычислении общий объем признака мысленно распределяется поровну между всеми единицами совокупности[[4]](#footnote-4).

Средняя арифметическая определяется по формулам:

1. Средней арифметической простой (для несгруппированных данных):

,(2.1)

где: – значение признака у i-й единицы совокупности; n – объем совокупности (Valid N).

1. Средней арифметической взвешенной (для интервального вариационного ряда):

, (2.2)

где:fi – абсолютные частоты; xi – середина интервала.

Определим среднюю арифметическую для рассматриваемых данных:

1. По формуле простой средней на основе массива несгруппированых данных:



1. По формуле средней арифметической взвешенной на основе группировочной таблицы с 8 интервалами (табл. 2.1):



Таблица 2.1 Расчет средней арифметической взвешенной для распределения регионов России по количеству легковых автомобилей на 1000 чел. населения за 2005 г.

|  |  |  |  |
| --- | --- | --- | --- |
| Интервал | Абсолютная частота (fi) | Середина интервала (xi) |  |
| 19,31429-49,68571 | 2 | 34,5 | 69 |
| 49,68571-80,05714 | 3 | 64,871425 | 194,6143 |
| 80,05714-110,4286 | 6 | 95,24287 | 571,4572 |
| 110,4286-140,8 | 15 | 125,6143 | 1884,215 |
| 140,8-171,1714 | 32 | 155,9857 | 4991,542 |
| 171,1714-201,5429 | 13 | 186,35715 | 2422,643 |
| 201,5429-231,9143 | 4 | 216,7286 | 866,9144 |
| 231,9143-262,2857 | 5 | 247,1 | 1235,5 |
| Итого:  | 80 | – | 12235,89 |

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменным произведение индивидуальных величин, то следует применить геометрическую среднюю величину. Ее формула такова:

 (2.3)

Значение средней геометрической было рассчитано с помощью ППП «Statistica» и составило 145,9133.

При изучении вариации применяются такие характеристики вариационного ряда, которые описывают количественно его структуру, строение. Такова, например, медиана – величина варьирующего признака, делящая совокупность на две равные части – со значениями признака меньше медианы и со значениями признака больше медианы.

В интервальном вариационном ряду для нахождения медианы применяется формула:

, (2.4)

где: Ме – медиана; Хе – нижняя граница интервала, в котором находится медиана; n – число наблюдений; fMe-1 – накопленная частота в интервале, предшествующем медианному; fMe – частота в медианном интервале; i – величина интервала.

Рассчитаем значение медианы вариационного ряда, использовав для этого таблицу распределения с 8-ю интервалами (табл. 2.1). Медианным интервалом является интервал 140,8-171,1714, следовательно нижняя граница медианного интервала – 140,8; величина интервала – 30,37 (164,42–171,17); кумулятивная частота предшествующего интервала – 26, частота медианного интервала – 32. Медиана вариационного ряда равна:



Значение медианы, рассчитанное с помощью программы Statistica (по исходному несгруппированному ряду данных), составляет 153,45.

Важное значение имеет такая величина признака, которая встречается в изучаемом ряду, в совокупности чаще всего. Такую величину принято называть модой и обозначать Мо. В дискретном ряду мода определяется без вычисления как значение признака с наибольшей частотой.

В интервальном вариационном ряду, тем более при непрерывной вариации признака, строго говоря, каждое значение признака встречается только один раз. Модальным интервалом является интервал с наибольшей частотой. Значение моды в интервальном ряду распределения определяется по следующей формуле:

, (2.5)

где:Х0 – нижняя частота модального интервала; fMo – частота в модальном интервале; fMo-1 – частота в предыдущем интервале; fMo+1 – частота в следующем интервале за модальным; i – величина интервала.

Модальным интервалом является интервал 140,8-171,17; нижняя граница интервала – 140,8; частота модального интервала – 32, частота предыдущего интервала – 15; частота следующего интервала – 13; величина интервала – 30,37.

Определим модальное значение:



В ППП Statistica значение моды определяется непосредственно по исходным несгруппированным данным. Для рассматриваемого случая модальное значение равно 161,7, а его частота составляет 2.

3. ОЦЕНКА ВАРИАЦИИ ИЗУЧАЕМОГО ПРИЗНАКА

Вариация – это различие в значениях какого-либо признака у разных единиц данной совокупности в один и тот же период или момент времени[[5]](#footnote-5).

Она возникает в результате того, что индивидуальные значения признака складываются под влиянием различных факторов, которые в разном случае могут сочетаться по-разному.

К показателям вариации относятся: размах вариации, дисперсия и среднее квадратическое отклонение, коэффициент вариации.

Простейшим показателем вариации является размах, или амплитуда вариации, – абсолютная разность между максимальным и минимальным значениями признака из имеющихся в изучаемой совокупности значений. Таким образом, размах вариации вычисляется по формуле:

, (3.1)

Минимальное значение признака (Xmin) для исследуемой совокупности составило 34,5, а максимальное (Xmax) – 247,1. Следовательно, размах вариации для вариационного ряда составляет:



Дисперсия признака представляет собой средний квадрат отклонений вариантов от их средней величины и вычисляется по следующим формулам:

Простая дисперсия для несгруппированных данных:

 (3.2)

Взвешенная дисперсия для вариационного ряда:

 (3.3)

Простая дисперсия по несгруппированным данным была рассчитана с помощью программы Statistica и составила 1730,257.

Взвешенная дисперсия по сгруппированным данным рассчитана в табл. 3.1

Таблица 3.1 Расчет дисперсии для распределения регионов России по количеству легковых автомобилей на 1000 чел. населения за 2005 г.

|  |  |  |  |
| --- | --- | --- | --- |
| Интервал | Абсолютная частота (fi) | Середина интервала (xi) |  |
| 19,31429-49,68571 | 2 | 34,50 | 28060,13 |
| 49,68571-80,05714 | 3 | 64,87 | 23272,75 |
| 80,05714-110,4286 | 6 | 95,24 | 19979,69 |
| 110,4286-140,8 | 15 | 125,61 | 11207,44 |
| 140,8-171,1714 | 32 | 155,99 | 295,17 |
| 171,1714-201,5429 | 13 | 186,36 | 14509,73 |
| 201,5429-231,9143 | 4 | 216,73 | 16271,57 |
| 231,9143-262,2857 | 5 | 247,10 | 44322,46 |
| Итого:  | 80 |  – | 157918,93 |

Взвешенная дисперсия вариационного ряда:



Среднее квадратическое отклонение – это обобщающая характеристика размеров вариации признака в совокупности. Оно показывает, на сколько в среднем отклоняются конкретные варианты от их среднего значения.

Среднее квадратическое отклонение является корню квадратному из дисперсии.

Определим среднее квадратическое отклонение:

* По исходному ряду данных:



* По сгруппированным данным:



Чем меньше значение дисперсии и среднего квадратического отклонения, тем более однородна исследуемая совокупность.

Для сравнения вариаций различных признаков используются относительные показатели вариации, в частности, коэффициент вариации.

Коэффициент вариации представляет собой выраженное в процентах отношение среднего квадратического отклонения к средней арифметической:

 (2.8)

Коэффициент вариации исследуемого ряда данных равен:



Коэффициент вариации используют не только для сравнительной оценки вариации единиц совокупности, но и как характеристику однородности совокупности. Совокупность считается количественно однородной, если коэффициент вариации не превышает 33%.

Так как коэффициент вариации в данном случае составляет 27,18%, т.е. значительно меньше 33%, то исследуемая совокупность является количественно однородной.

4. ХАРАКТЕРИСТИКА СТРУКТУРЫ РАСПРЕДЕЛЕНИЯ

Структура распределения характеризуется такими показателями, как медиана, квартили и децили. Медиана ряда распределения была определена в разделе 2, она составила 153,45.

Аналогично медиане вычисляются значения признака, делящие совокупность на четыре равные по числу единиц части. Эти величины называются квартилями и обозначаются заглавной латинской буквой Q с подписным значком номера квартиля. Очевидно, что Q2 равно медиане распределения.

Значения признака, делящие ряд на пять равных частей, называют квинтилями, на десять частей – децилями, на сто частей – перцентилями.

Для расчёта квартилей применяются следующие формулы:

1. для несгруппированных данных:

Нижний (первый) квартиль (Lower quartile):

, ,  (4.1)

Верхний (третий) квартиль (Upper quartile):

, ,  (4.2)

1. в интервальном вариационном ряду распределения:

 (4.3)

 (4.4)

где:Q1 и Q3 – нижний и верхний квартили; ,  – нижние границы квартильных интервалов; h – величина группировочного интервала;  – абсолютные частоты квартильных интервалов;  – накопленные (кумулятивные) частоты интервалов, предшествующих квартильным.

Рассчитаем квартили распределения на основе сгруппированных данных (табл. 4.1).

Таблица 4.1 Исходные данные для расчета квартилей распределения регионов России по количеству легковых автомобилей на 1000 чел. населения за 2005 г.

|  |  |  |
| --- | --- | --- |
| Интервал | Абсолютная частота (fi) | Кумулятивная частота (Fi) |
| 19,31429-49,68571 | 2 | 2 |
| 49,68571-80,05714 | 3 | 5 |
| 80,05714-110,4286 | 6 | 11 |
| 110,4286-140,8 | 15 | 26 |
| 140,8-171,1714 | 32 | 58 |
| 171,1714-201,5429 | 13 | 71 |
| 201,5429-231,9143 | 4 | 75 |
| 231,9143-262,2857 | 5 | 80 |
| Итого:  | 80 |  – |

Нижний квартиль распределения равен:



Верхний квартиль распределения:



Квартили, рассчитанные с помощью программы Statistica, немного отличаются от тех, что рассчитаны вручную по сгруппированным данным:

* Нижний квартиль равен 135,85.
* Верхний квартиль – 172,75.

5. ХАРАКТЕРИСТИКА ФОРМЫ РАСПРЕДЕЛЕНИЯ

Для дальнейшего изучения характера вариации используются средние значения разных степеней отклонений отдельных величин признака от его средней арифметической величины.

Эти показатели получили название центральных моментов распределения порядка, соответствующего степени, в которую возводятся отклонения (табл. 5.1), или просто моментов.

Таблица 5.1 Формулы для расчета центральных моментов

|  |  |
| --- | --- |
| Порядок момента | Формулы для расчета |
| для несгруппированных данных | для сгруппированных данных |
| Первый  |  |  |
| Второй  |  |  |
| Третий  |  |  |
| Четвертый  |  |  |

Согласно свойству средней арифметической центральный момент первого порядка равен нулю, второй центральный момент представляет собой дисперсию. Третий центральный момент используется для оценки асимметрии распределения, четвертый – для оценки эксцесса.

На основе момента третьего порядка можно построить показатель, характеризующий степень асимметричности распределения:

 (5.1)

Этот показатель называют коэффициентом асимметрии. Он может быть рассчитан как по сгруппированным, так и по несгруппированным данным.

С помощью момента четвертого порядка характеризуется свойство рядов распределения, называемое эксцессом. Показатель эксцесса рассчитывается по формуле:

 (5.2)

Исходные данные для расчета асимметрии и эксцесса приведены в табл. 5.2.

Таблица 5.2 Расчет ассиметрии и эксцесса для распределения регионов России по количеству легковых автомобилей на 1000 чел. населения за 2005 г.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Интервал | Абсолютная частота (fi) | Середина интервала (xi) |  |  |
| 19,31429-49,68571 | 1 | 34,5 | -3323682,13 | 393685400,9 |
| 49,68571-80,05714 | 3 | 58,12 | -2049797,54 | 180540318,8 |
| 80,05714-110,4286 | 3 | 81,74 | -1152941,93 | 66531323,0 |
| 110,4286-140,8 | 5 | 105,365 | -306347,11 | 8373775,1 |
| 140,8-171,1714 | 14 | 128,99 | 896,48 | 2722,7 |
| 171,1714-201,5429 | 28 | 152,61 | 484749,46 | 16194790,2 |
| 201,5429-231,9143 | 14 | 176,23 | 1037801,06 | 66190981,2 |
| 231,9143-262,2857 | 6 | 199,855 | 4173022,64 | 392896041,5 |
| Итого:  | 80 |  – | -1136299,06 | 1124415353,4 |

Коэффициент ассиметрии по сгруппированным данным:



Коэффициент ассиметрии на основе исходного ряда данных был рассчитан с помощью ППП Statistica и составил -0,341.

Коэффициент эксцесса на основе сгруппированных данных:



Коэффициент эксцесса, рассчитанный для несгруппированных данных, составил 1,075.

Сопоставим показатели, рассчитанные вручную по сгруппированным данным, и показатели, полученные с помощью программы Statistica на основе исходного ряда данных (табл. 5.3).

Таблица 5.3 Сравнение статистических показателей, рассчитанных различными способами

|  |  |  |  |
| --- | --- | --- | --- |
| № | Название показателя | Значение в ППП Statistica | Значение после ручного расчета |
|  | Средняя арифметическая | 153,055 | 152,95 |
|  | Медиана | 153,45 | 154,09 |
|  | Мода | 161,70 | 155,14 |
|  | Дисперсия | 1730,257 | 1973,99 |
|  | Нижний квартиль | 135,85 | 128,65 |
|  | Верхний квартиль | 172,75 | 175,84 |

6. СГЛАЖИВАНИЕ ЭМПИРИЧЕСКОГО РАСПРЕДЕЛЕНИЯ.

ПРОВЕРКА ГИПОТЕЗЫ О ЗАКОНЕ РАСПРЕДЕЛЕНИЯ

Одна из важнейших задач анализа вариационных рядов заключается в выявлении закономерности распределения и определении ее характера. Основной путь в выявлении закономерности распределения – построение вариационных рядов для достаточно больших совокупностей. Важное значение для выявления закономерности распределения имеет правильное построение самого вариационного ряда: выбор числа групп и размера интервала варьирующего признака.

Говоря о характере, типе закономерности распределения, имеем в виду отражение в нем общих условий вариации. При этом речь всегда идет о распределениях качественно однородных явлений. Общие условия, определяющие тип закономерности распределения, познаются анализом сущности явления, тех его свойств, которые определяют вариацию изучаемого признака. Следовательно, должна быть выдвинута какая-то научная гипотеза, обосновывающая тип теоретической кривой распределения.

Под теоретической кривой распределения понимается графическое изображение ряда в виде непрерывной линии изменения частот в вариационном ряду, функционально связанного с изменением вариантов (значений признака).

Теоретическое распределение может быть выражено аналитически – формулой, которая связывает частоты вариационного ряда и соответствующие значения признака. Такие алгебраические формулы носят название законов распределения[[6]](#footnote-6).

Процедура выравнивания, сглаживания анализируемого распределения заключается в замене эмпирических частот теоретическими, определяемыми по формуле теоретического распределения, но с учетом фактических значений переменной. На основе сопоставления эмпирических и теоретических частот рассчитываются критерии согласия, которые используются для проверки гипотезы о соответствии исследуемого распределения тому или иному типу теоретического распределении.

Для проверки статистической гипотезы о законе распределения будем использовать критерий  – критерий Пирсона (Chi-square test). Расчет критерия производится по следующей формуле:

 (6.1)

где: – эмпирические абсолютные частоты (Observed Frequency);  – абсолютные частоты теоретического распределения (Expected Frequency); k – число интервалов.

С помощью ППП Statistica проведем сглаживание рассматриваемого распределения и проверим статистическую гипотезу о законе распределения.

Рис. 6.1. Проверка гипотезы о нормальном распределении переменной Var1

Для сглаживания эмпирического распределения переменной Var1 нормальным распределением необходимо использовать формулы, приведенные ниже.

Функция нормального распределения:

 (6.2)

Плотность нормального распределения определяется по формуле:

 (6.3)

где:х – значение изучаемого признака;  – средняя арифметическая величина;  – среднее квадратическое отклонение изучаемого признака;  – математические константы;  – нормированное отклонение.

Теоретические частоты нормального распределения рассчитываются по следующей формуле:

 (6.4)

где:N – объем совокупности; h – величина интервала.

Из рис. 6.1. видно, что критерий  для нормального распределения составил 5,42808 при количестве степеней свободы 2 и расчетном уровне значимости 0,06627.

Для принятия решения о справедливости гипотезы о законе распределения необходимо сравнить рассчитанный критерий  с критическим значением.

Табличное значение  для степеней свободы r=2 и уровня значимости α=0,05 составляет 5,991. Поскольку рассчитанное значение  меньше табличного, то гипотеза о нормальном распределении переменной Var1 не противоречит статистическим данным.

На рис. 6.2 показана гистограмма эмпирического распределения и расчетная кривая нормального распределения для исследуемой переменной.

Рис. 6.2. Гистограмма и расчетная кривая нормального распределения для переменной Var1

В табл. 6.1 приведен расчет теоретических частот для сглаживания эмпирических данных нормальным распределением.

Расчетное значение критерия Пирсона составило . Табличное значение критерия – .

Таблица 6.1 Расчет критерия  вручную

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| № |  |  |  |  |  |  |  |
|  | 19,31 | 49,69 | 34,50 | -2,666 | 0,0114 | 1 | 1,000 |
|  | 49,69 | 80,06 | 64,87 | -1,982 | 0,0559 | 3 | 0,000 |
|  | 80,06 | 110,43 | 95,24 | -1,299 | 0,1716 | 9 | 1,000 |
|  | 110,43 | 140,80 | 125,61 | -0,615 | 0,3302 | 18 | 0,500 |
|  | 140,80 | 171,17 | 155,99 | 0,068 | 0,3980 | 23 | 3,522 |
|  | 171,17 | 201,54 | 186,36 | 0,752 | 0,3007 | 16 | 0,563 |
|  | 201,54 | 231,91 | 216,73 | 1,436 | 0,1424 | 8 | 2,000 |
|  | 231,91 | 262,29 | 247,10 | 2,119 | 0,0422 | 2 | 4,500 |
| Итого: | 13,084 |

Очевидно, что расчетное значение критерия превышает критическое, следовательно гипотеза о нормальном распределении подтверждена (табл. 6.2).

Таблица 6.2 Проверка гипотезы о нормальном законе распределения вручную

|  |  |  |  |
| --- | --- | --- | --- |
| Тип распределения | Число степеней свободы r | Расчетное значение критерия  | Табличное значение критерия  |
| Нормальное | 7 | 13,084 | 14,07 |

Рассмотрим также гипотезы о логнормальном и прямоугольном распределении (рис. 6.2 и рис. 6.3).

Из рис. 6.2 видно, что критерий  для логнормального распределения равен 16,48145 при количестве степеней свободы r=3 и уровне значимости 0,0009.

Рис. 6.2. Проверка гипотезы о логарифмически нормальном распределении переменной Var1

Сопоставим рассчитанные показатели с табличным значением критерия Пирсона:





Очевидно, что расчетное значение критерия Пирсона превышает критическое, а расчетная вероятность ниже табличного уровня значимости. Следовательно, гипотеза о логнормальном распределении вариационного ряда не может быть принята.

На рис. 6.3 приведена гистограмма и расчетная кривая логнормального распределения переменной Var1.

На рис. 6.4 приведена таблица расчета теоретических частот и критерия Пирсона для прямоугольного распределения.

Таким образом, расчетный критерий Пирсона для прямоугольного распределения составил 54,48687 при количестве степеней свободы 5 и вероятности 0,00:





Рис. 6.3. Гистограмма и расчетная кривая логнормального распределения для переменной Var1

Рис. 6.4. Проверка гипотезы о прямоугольном распределении переменной Var1

Очевидно, что расчетный критерий Пирсона  намного превышает табличное значение, следовательно гипотеза о прямоугольном распределении переменной Var1 отклонена.

ЗАКЛЮЧЕНИЕ

В курсовой работе были проанализированы данные о распределении регионов России по количеству легковых автомобилей на 1000 человек населения за 2005 г. Для удобства анализа данные были представлены в виде группировочных таблиц с количеством интервалов n=8, 10 и 13. Наиболее пригодной для анализа оказалась группировочная таблица с восемью интервалами.

Также для удобства анализа вариационного ряда используется графическое представление. В работе были использованы такие виды графиков, как полигон, кумулята и гистограмма. Полигон, построенный на основе абсолютных частот, показывает форму распределения. Из рисунка видно, что распределение имеет одну вершину, форма его симметрична и довольно крута.

Также с помощью графика можно определить модальный интервал (140,8–171,17). Гистограмма позволяет сделать такие же выводы.

Кумулята показывает накопленные частоты распределения (абсолютные или относительные). С помощью кумуляты легко определить медианный интервал распределения (140,8–171,17) – это интервал, на котором кумулята переваливает за середину распределения, т.е. за 40 (для абсолютных частот) или 50% (для относительных частот). Так как модальный и медианный интервалы распределения совпадают, то распределение симметрично.

Центральная тенденция распределения характеризуется такими показателями, как среднее арифметическое значение, мода и медиана. Все показатели были определены с помощью программы Statistica по исходному ряду данных и вручную по сгруппированным данным. Среднее арифметическое значение вариационного ряда составило 153,055 (по исходным данным) и 152,95 (по группировочной таблице).

Медиана – это величина признака, делящая распределение на две равные части. По исходным данным медиана составила 153,45, а по сгруппированным данным – 154,09.

Мода – это значение признака с наибольшей частотой. Ее значение составило 155,14. Очевидно, что и среднее арифметическое, и медиана, и мода принадлежат одному интервалу и незначительно отличаются по значениям. Это свидетельствует о симметричности распределения относительно центра.

Вариация – это различие в значениях какого-либо признака у разных единиц данной совокупности в один и тот же период или момент времени. К показателям, характеризующим вариацию распределения, относятся размах вариации, дисперсия и среднее квадратическое отклонение и коэффициент вариации.

Размах вариации показывает амплитуду вариации и определяется как разница между максимальным и минимальным значением распределения и составляет 212,6.

Дисперсия признака представляет собой средний квадрат отклонений вариантов от их средней величины. Дисперсия, рассчитанная по исходным данным, составила 1730,257, а по сгруппированным – 1973,99. Более удобным для анализа показателем является среднее квадратическое отклонение, которое определяется как корень из дисперсии. Среднее квадратическое отклонение, рассчитанное на основании исходного ряда распределения, равно 41,596, а отклонение, определенное по сгруппированным данным, – 44,43. Оно показывает, что значение признака отклоняется от среднего арифметического значения в среднем на 41,596.

Коэффициент вариации определяется как отношение среднего квадратического отклонения к среднему арифметическому значению. Этот показатель используют для характеристики однородности совокупности. Значение коэффициента вариации для исследуемого ряда данных составило 27,18%. Поскольку рассчитанное значение коэффициента меньше 33%, то данная совокупность является количественно однородной.

Структура распределения характеризуется такими показателями, как медиана, квартили и децили. Медиана делит совокупность на две равные части, квартили – на четыре части, а децили – на 10 частей. Медиана распределения составляет 153,45, нижний квартиль – 135,85, верхний квартиль – 172,75. Разница между первым и вторым квартилем (медианой) составляет 17,6; между вторым и третьим – 19,3. Очевидно, что квартили расположены очень близко один к другому, что говорит о высокой плотности середины распределения.

Форма распределения характеризуется асимметрией и эксцессом. Коэффициент асимметрии показывает, как следует из названия, степень асимметричности распределения и определяется как отношение третьего центрального момента к стандартному отклонению в кубе. Коэффициент асимметрии исследуемого распределения равен -0,341 (по сгруппированным данным – -0,168), что свидетельствует о небольшой левосторонней асимметрии.

Эксцесс характеризует «крутизну» распределения и определяется как отношение четвертого центрального момента к стандартному отклонению в четвертой степени. Для нормального распределения величина эксцесса равна трем, поэтому от рассчитанного значения отнимают 3. Значение эксцесса для анализируемого распределения равно 1,075 (рассчитанное вручную по группировочной таблице – 0,824). Это означает, что исследуемое распределение гораздо «круче» нормального.

Одна из важнейших задач анализа вариационных рядов заключается в выявлении закономерности распределения и определении ее характера. Для этого осуществляется процедура выравнивания и проверка гипотезы о соответствии эмпирического ряда данных теоретическому распределению. В данной работе были проверены гипотезы соответствия эмпирического вариационного ряда нормальному, логнормальному и прямоугольному распределениям с помощью критерия Пирсона. Для этого с помощью программы Statistica было осуществлено сглаживание эмпирического ряда данных путем расчета теоретических частот и сравнение полученных значений с эмпирическими частотами. В результате этих расчетов было получено расчетное значение критерия .

Расчетный критерий для нормального распределения составил 5,42808 при количестве степеней свободы 2 и расчетном уровне значимости 0,06627. Табличное значение критерия равно 5,991. Поскольку расчетное значение меньше критического, гипотеза о нормальном распределении не противоречит статистическим данным.

Аналогично были рассчитаны критерии Пирсона для логнормального () и прямоугольного () распределения. Оба критерия значительно превышают соответствующие табличные значения, следовательно, гипотезы о логнормальном и прямоугольном распределениях не подтвердились.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Боровиков В.П., STATISTICA. Искусство анализа данных на компьютере: для профессионалов / В. П. Боровиков. – 2-е изд. – СПб. : – 2003. – 688 с.
2. Венецкий И.Г., Основные математико-статистические понятия и формулы в экономическом анализе. Справочник / И.Г. Венецкий, В.И. Венецкая. – 2-е изд., перераб. и доп. – М. : Статистика, 1979. – 477 с.
3. Гмурнан В.Э. Теория вероятностей и математическая статистика: Учеб. пособие. – М.: Высш. шк., 2003. – 479 с.
4. Гусаров В.М. Статистика: Учеб. пособие для вузов. – М.: ЮНИТИ-ДАНА, 2003. – 463 с.
5. Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник / Под ред. И.И. Елисеевой. – 5-е изд., перераб. и доп. – М.: Финансы и статистика, 2004. – 656 с.
6. Закс Л., Статистическое оценивание: Пер. с нем / Л. Закс. – М.: Статистика, 1976. – 597 с.
7. Н.В. Куприенко Статистика. Методы анализа распределений. Выборочное наблюдение. 3-е изд. : учеб. пособие. / Н.В. Куприенко, О.А. Пономарева, Д.В. Тихонов. – СПб.: Изд-во Политехн. ун-та, 2009. – 138 с.
8. Общая теория статистики: Статистическая методология в изучении коммерческой деятельности: Учебник. / Под ред. А.А. Спирина, О.Э. Башиной. – М.: Финансы и статистика, 1996. – 296 с.
9. Общая теория статистики: учеб. / М.Р. Ефимова, Е.В. Петрова, В.Н. Румянцев. – М.: ИНФРА-М, 2002. – 416 с.
10. Орлов А.И. Прикладная статистика. Учебник. / А.И. Орлов. – М.: Издательство «Экзамен», 2004. – 656 с.
11. Регионы России. Социально-экономические показатели. 2006: Стат.сб. М., 2007.
12. Сизова Т.М. Статистика: Учебное пособие. – СПб.: СПб ГУИТМО, 2005. – 80 с.
13. Теория статистики.: учеб. /Под ред. Р.А. Шмойловой. – М.: Финансы и статистика, 2005. – 560 с.
14. Теория статистики: учеб. / Под ред. проф. Г.Л. Громыко. – 2-е изд., перераб. и доп. – М.: ИНФРА-М, 2006. – 476 с.
15. Экономическая статистика: Учебник. / Под ред. Ю.Н. Иванова. – М.: ИНФРА-М, 2004. – 480 с.
1. Гусаров В.М. Статистика: Учеб. пособие для вузов. – М.: ЮНИТИ-ДАНА, 2003. – С. 43 [↑](#footnote-ref-1)
2. Гусаров В.М. Статистика: Учеб. пособие для вузов. – М.: ЮНИТИ-ДАНА, 2003. – С. 39 [↑](#footnote-ref-2)
3. Общая теория статистики: Статистическая методология в изучении коммерческой деятельности: Учебник. / Под ред. А.А. Спирина, О.Э. Башиной. – М.: Финансы и статистика, 1996. – С. 62 [↑](#footnote-ref-3)
4. Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник / Под ред. И.И. Елисеевой. – 5-е изд., перераб. и доп. – М.: Финансы и статистика, 2004. – С. 124 [↑](#footnote-ref-4)
5. Гусаров В.М. Статистика: Учеб. пособие для вузов. – М.: ЮНИТИ-ДАНА, 2003. – С. 71 [↑](#footnote-ref-5)
6. Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник / Под ред. И.И. Елисеевой. – 5-е изд., перераб. и доп. – М.: Финансы и статистика, 2004. – С. 274 [↑](#footnote-ref-6)