АНАЛИЗ РЯДОВ РАСПРЕДЕЛЕНИЯ

Содержание

Введение

1. Характеристики центра распределения

1.1 Мода

1.2 Медиана

1.3 Показатели дифференциации

2. Характеристики вариации

2.1 Абсолютные характеристики вариации

2.1.1 Расчет дисперсии способом моментов

2.1.2 Расчет дисперсии альтернативного признака

2.1.3 Межгрупповая дисперсия. Правило сложения дисперсий

2.2 Относительные характеристики вариации

3. Теоретические кривые распределения

3.1 Нормальное распределение

3.2 Выравнивание эмпирического распределения по кривой нормального распределения

3.3 Критерии согласия

3.4 Характеристики неравномерности распределения

## Введение

Ряд распределения (т.е. упорядоченное распределение единиц изучаемой совокупности на группы по определенному варьирующему признаку) характеризует состав, структуру совокупности по определенному признаку. Его строят для того, чтобы выявить характер распределения единиц совокупности по варьирующему признаку, определить закономерности в этом распределении.

Для анализа ряда распределения используют ряд статистических характеристик:

частотные характеристики;

характеристики центра распределения;

характеристики вариации;

характеристики неравномерности распределения.

Частотные характеристики ряда распределения, а именно, частоты и частости (или другое название - доля ), накопленные (или кумулятивные) частоты и частости , абсолютная и относительная плотность распределения, были рассмотрены в теме "Сводка и группировка статистических данных".


## 1. Характеристики центра распределения

К характеристикам центра распределенияотносят среднюю, моду и медиану. Эти характеристики принято также называть структурными средними, они определяют вид полигона и гистограммы, эмпирического закона распределения**.**

В качестве средней для характеристики центра распределениячаще всего используют среднюю арифметическую простую или взвешенную.

## 1.1 Мода

Мода (Мо) - это варианта, которая чаще всего встречается в изучаемой совокупности. Мода не зависит от крайних значений вариант и может применяется для характеристики центра в рядах распределения с неопределенными границами.

В дискретном вариационном ряду мода определяется визуально и равна варианте с наибольшей частотой или частостью. Данные распределения рабочих по стажу работы (см. лекцию "Сводка и группировка статистических данных") показывают, что наибольшее рабочих имеют стаж работы 4 года, т.е. варианта, равная 4, является модой признака. Мо = 4.

В интервальных рядах распределения для нахождения моды сначала по наибольшей частоте определяют модальный интервал, т.е. интервал, содержащий моду, а затем приблизительно рассчитывают ее по формуле:

,

где - нижняя граница модального интервала;

- величина модального интервала;

 - частоты соответственно в предыдущем и следующим за модальным интервалах.

Встречаются ряды, которые имеют две моды (бимодальный ряд) или несколько (полимодальный).

Рассчитаем моду интервального ряда распределения рабочих по размеру заработной платы (см. лекцию "Сводка и группировка статистических данных").

В этом вариационном ряду интервал 900-1000 грн., в который попало максимальное количество рабочих (9 чел), является модальным.

 грн.

Полученное значение моды свидетельствует о том, что в рассматриваемой совокупности наиболее типичной является заработная плата 914,29 грн., что выше ранее рассчитанной средней зарплаты (870 грн).

Для ряда с неравными интервалами модальный интервал определяется по наибольшей плотности распределения, а в расчетной формуле моды вместо частот используют абсолютные плотности распределения.

Для интервальных вариационных рядов с равными интервалами моду можно приближенно определить графически.

Для этого на гистограмме этого ряда (см. гистограмму в лекции "Сводка и группировка статистических данных") выбирают самый высокий прямоугольник, который и является модальным.

Далее правую верхнюю вершину прямоугольника, предшествующего модальному (частота fMо-1), соединяют с правой верхней вершиной модального прямоугольника (частота fMо), а левую верхнюю вершину этого прямоугольника - с левой верхней вершиной прямоугольника, следующего за модальным (частота fMо+1).

Из точки пересечения опускают перпендикуляр на горизонтальную ось. Основание перпендикуляра покажет значение моды Мо. Точность определения зависит от масштаба графика.

## 1.2 Медиана

Медианой Ме называют такое значение признака, которое приходится на середину ранжированного ряда и делит его на две равные по числу единиц части. Таким образом, в ранжированном ряду распределения одна половина ряда имеет значения признака, превышающие медиану, другая - меньше медианы. Медиану используют вместо средней арифметической, когда крайние варианты ранжированного ряда (наименьшая и наибольшая) по сравнению с остальными оказываются чрезмерно большими или чрезмерно малыми.

В дискретном вариационном ряду, содержащем нечетное число единиц, медиана равна варианте признака, имеющей номер

:

,

где N - число единиц совокупности.

В дискретном ряду, состоящем из четного числа единиц совокупности, медиана определяется как средняя из вариант, имеющих номера

 и : .

В распределении рабочих по стажу работы медиана равна средней из вариант, имеющих в ранжированном ряду номера 10: 2 = 5 и 10: 2 + 1 = 6. Варианты пятого и шестого признака равны 4 годам, таким образом

года

При вычислении медианы в интервальном ряду сначала находят медианный интервал, (т.е. содержащий медиану), для чего используют накопленные частоты или частости. Медианным является интервал, накопленная частота которого равна или превышает половину всего объема совокупности. Затем значение медианы рассчитывается по формуле:

,

где - нижняя граница медианного интервала; - ширина медианного интервала; - накопленная частота интервала, предшествующего медианному; - частота медианного интервала.

Рассчитаем медиану ряда распределения рабочих по размеру зарплаты (см. лекцию "Сводка и группировка статистических данных").

Медианным является интервал заработной платы 800-900 грн., поскольку его кумулятивная частота равна 17, что превышает половину суммы всех частот (). Тогда

Ме=800+100грн.

Полученное значение говорит о том, половина рабочих имеют заработную плату ниже 875 грн., но это выше среднего ее размера.

Для определения медианы можно вместо кумулятивных частот использовать кумулятивные частости .

Медиана, как и мода, не зависит от крайних значений вариант, поэтому также применяется для характеристики центра в рядах распределения с неопределенными границами.

**Свойство медианы**: сумма абсолютных величин отклонений вариант от медианы меньше, чем от любой другой величины (в том числе и от средней арифметической):

Это свойство медианы используется на транспорте при проектировании расположения трамвайных и троллейбусных остановок, бензоколонок, сборочных пунктов и т. д.

**Пример.** На шоссе длиной 100 км расположено 10 гаражей. Для проектирования строительства бензоколонки были собраны данные о числе предполагаемых ездок на заправку по каждому гаражу.

Таблица 2 - Данные о количестве ездок на заправку по каждому гаражу.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Километр шоссе, на котором расположен гараж | 7 | 26 | 28 | 37 | 40 | 46 | 60 | 78 | 86 | 92 | Всего ездок |
| Проектируемое число ездок | 10 | 15 | 5 | 20 | 5 | 25 | 15 | 30 | 10 | 65 | 200 |

Нужно поставить бензоколонку так, чтобы общий пробег автомашин на заправку был наименьшим.

Вариант 1. Если бензоколонку поставить в середине шоссе, т.е. на 50-ом километре (центр диапазона изменения признака), то пробеги с учетом числа ездок составят:

а) в одном направлении:

;

б) в противоположном:

;

в) общий пробег в оба направления: .

Вариант 2. Если бензоколонку поставить на среднем участке шоссе, определенном по формуле средней арифметической с учетом числа ездок:

Тогда пробеги составят:

а) в одном направлении:

б) в противоположном:

;

в) общий пробег в оба направления, равный меньше, чем в первом варианте на 438,5 км.

Вариант 3. Если поставить бензоколонку на 78-м километре, что будет соответствовать медиане по количеству ездок (накопленное число ездок для 60 км - 95, для 78 км - 125).

Тогда пробеги составят:

а) в одном направлении:

б) в противоположном:

;

в) общий пробег: , меньше общих пробегов, рассчитанных по предыдущим вариантам.

Таким образом, медиане соответствует наилучший результат, т.е. минимальный общий пробег.

Медиану можно определить графически, по кумуляте (см. лекцию "Сводка и группировка статистических данных"). Для этого последнюю ординату, равную сумме всех частот или частостей, делят пополам. Из полученной точки восстанавливают перпендикуляр до пересечения с кумулятой. Абсцисса точки пересечения и дает значение медианы.

## 1.3 Показатели дифференциации

Если возникает необходимость изучить структуру вариационного ряда более подробно, вычисляют значения признака, аналогичные медиане. Такие значения признака, которые делят все единицы распределения на равные численности, называют квантилями, или градиентами. Квартили и децили - частные случаи квантилей.

Квартилями (Q) называют значения признака, которые делят совокупность на четыре равные по числу единиц части. Децили (D) - признаки, делящие совокупность на десять равных частей.

Следовательно, кроме медианы, в ряду распределения имеются три квартиля и девять децилей. Медиана одновременно является вторым квартилем и пятым децилем. Расчет первого (Q1) и третьего (Q3) квартилей аналогичен расчету медианы, только вместо медианного интервала берется для первого квартиля интервал, в котором находится варианта, отсекающая ¼ численности частот, а для третьего квартиля - ¾ численности частот:

 и .

Логика построения квинтилей и децилей аналогична.

## 2. Характеристики вариации

Показатели вариации характеризует колеблемость индивидуальных значений признака по отношению к среднему значению, что не менее важно, чем определение самой средней. Средняя не показывает строения совокупности, как располагаются около нее варианты осредняемого признака, сосредоточены ли они вблизи средней или значительно отклоняются от нее. Средняя величина признака в двух совокупностях может быть одинаковой, но в одном случае все индивидуальные значения отличаются от нее мало, а в другом эти отличия велики, т.е. в одном случае вариация признака мала, а в другом велика.

Это можно показать на таком примере. Предположим, что две бригады из 3-х человек каждая выполняют одинаковую работу. Количество деталей, изготовленных за смену отдельными рабочими, составило:

в первой бригаде - 95, 100, 105;

во второй бригаде - 75, 100, 125.

Средняя выработка на одного рабочего в бригадах составила

, .

Средняя выработка одинакова, но колеблемость выработки отдельных рабочих в первой бригаде значительно меньше, чем во второй.

Следовательно, чем больше варианты отдельных единиц совокупности различаются между собой, тем больше они отличаются от своей средней, и наоборот - варианты, мало отличающиеся друг от друга, более близки по значению к средней, которая в таком случае будет более реально представлять всю совокупность.

Поэтому для характеристики и измерения вариации признака в совокупности кроме средней используют следующие показатели:

абсолютные - вариационный размах, среднее линейное и среднее квадратическое отклонение, дисперсию;

относительные - коэффициенты вариации.

## 2.1 Абсолютные характеристики вариации

Вариационный размах (или размах вариации) - это разница между максимальным и минимальным значениями признака:

В нашем примере размах вариации сменной выработки рабочих составляет: в первой бригаде R=105-95=10 дет., во второй бригаде R=125-75=50 дет. (в 5 раз больше). Это говорит о том, что выработка 1-й бригады более "устойчива", но резервов роста выработки больше у второй бригады, т.к в случае достижения всеми рабочими максимальной для этой бригады выработки, ею может быть изготовлено 3\*125=375 деталей, а в 1-й бригаде только 105\*3=315 деталей.

Если крайние значения признака не типичны для совокупности, то используют квартильный или децильный размахи. Квартильный размах RQ= Q3-Q1 охватывает 50% объема совокупности, децильный размах первый RD1 = D9-D1охватывает 80% данных, второй децильный размах RD2= D8-D2 - 60%.

Недостатком показателя вариационного размаха является, но что его величина не отражает все колебания признака.

Простейшим обобщающим показателем, отражающим все колебания признака, является среднее линейное отклонение, представляющее собой среднюю арифметическую абсолютных отклонений отдельных вариант от их средней величины: для несгруппированных данных

,

для сгруппированных данных

,

где хi - значение признака в дискретном ряду или середина интервала в интервальном распределении.

В вышеприведенных формулах разности в числителе взяты по модулю, иначе, согласно свойству средней арифметической, числитель всегда будет равен нулю. Поэтому среднее линейное отклонение в статистической практике применяют редко, только в тех случаях, когда суммирование показателей без учета знака имеет экономический смысл. С его помощью, например, анализируется состав работающих, рентабельность производства, оборот внешней торговли.

Дисперсия признака - это средний квадрат отклонений вариант от их средней величины:

простая дисперсия

,

взвешенная дисперсия

.

Формулу для расчета дисперсии можно упростить:

Таким образом, дисперсия равна разности средней из квадратов вариант и квадрата средней из вариант совокупности:

.

Однако, вследствие суммирования квадратов отклонений дисперсия дает искаженное представление об отклонениях, поэтому ее на основе рассчитывают среднее квадратическое отклонение, которое показывает, на сколько в среднем отклоняются конкретные варианты признака от их среднего значения. Вычисляется путем извлечения квадратного корня из дисперсии:

для несгруппированных данных

,

для вариационного ряда

Чем меньше значение дисперсии и среднего квадратического отклонения, тем однороднее совокупность, тем более надежной (типичной) будет средняя величина.

Среднее линейное и среднее квадратичное отклонение - именованные числа, т.е. выражаются в единицах измерения признака, идентичны по содержанию и близки по значению. Рассчитывать абсолютные показатели вариации рекомендуется с помощью таблиц.

Таблица 3 - Расчет характеристик вариации (на примере срока данных о сменной выработке рабочих бригады)

|  |  |  |  |
| --- | --- | --- | --- |
| Группы рабочих по выработке, шт.  | Число рабочих,  | Середина интервала,  | ***Расчетные значения*** |
|  |  |  |  |  |
| 170-190 | 10 | 180 | 1800 | -36 | 360 | 1296 | 12960 |
| 190-210 | 20 | 200 | 4000 | -16 | 320 | 256 | 5120 |
| 210-230 | 50 | 220 | 11000 | 4 | 200 | 16 | 800 |
| 230-250 | 20 | 240 | 4800 | 24 | 480 | 576 | 11520 |
| **Итого:**  | 100 | - | 21600 | - | 1360 | - | 30400 |

Среднесменная выработка рабочих:

Среднее линейное отклонение:

Дисперсия выработки:

Среднее квадратическое отклонение выработки отдельных рабочих от средней выработки:

.


## 2.1.1 Расчет дисперсии способом моментов

Вычисление дисперсий связано с громоздкими расчетами (особенно если средняя величина выражена большим числом с несколькими десятичными знаками). Расчеты можно упростить, если использовать упрощенную формулу и свойства дисперсии.

Дисперсия обладает следующими свойствами:

если все значения признака уменьшить или увеличить на одну и ту же величину А, то дисперсия от этого не уменьшится:

,

если все значения признака уменьшить или увеличить в одно и то же число раз (h раз), то дисперсия соответственно уменьшится или увеличится в раз.

То есть, если дисперсию уменьшенных значений признака описать следующим выражением

, то или

Используя свойства дисперсии и сначала уменьшив все варианты совокупности на величину А, а затем разделив на величину интервала h, получим формулу вычисления дисперсии в вариационных рядах с равными интервалами способом моментов:

**,**

где - дисперсия, исчисленная по способу моментов;

h - величина интервала вариационного ряда;

 - новые (преобразованные) значения вариант;

А- постоянная величина, в качестве которой используют середину интервала, обладающего наибольшей частотой; либо вариант, имеющий наибольшую частоту;

- квадрат момента первого порядка;

 - момент второго порядка.

Выполним расчет дисперсии способом моментов на основе данных о сменной выработке рабочих бригады.

Таблица 4 - Расчет дисперсии по способу моментов

|  |  |  |  |
| --- | --- | --- | --- |
| Группы рабочих по выработке, шт.  | Число рабочих,  | Середина интервала,  | ***Расчетные значения*** |
|  |  |  |
| 170-190 | 10 | 180 | -2 | -20 | 40 |
| 190-210 | 20 | 200 | -1 | -20 | 20 |
| 210-230 | 50 | 220 | 0 | 0 | 0 |
| 230-250 | 20 | 240 | 1 | 20 | 20 |
| Итого | **100** | **-** | **-** | **-20** | **80** |

Порядок расчета:

определяем постоянное число А, это варианта с наибольшей частотой: А=220;

определяем ;

рассчитываем и ;

определяем моменты 1-го и 2-го порядка:

рассчитываем дисперсию:


## 2.1.2 Расчет дисперсии альтернативного признака

Среди признаков, изучаемых статистикой, есть и такие, которым свойственны лишь два взаимно исключающих значения.

Это альтернативные признаки.

Им придается соответственно два количественных значения: варианты 1 и 0.

Частостью варианты 1, которая обозначается p, является доля единиц, обладающих данным признаком. Разность 1-р=q является частостью варианты 0. Таким образом,

|  |  |
| --- | --- |
| хi | wi |
| 1 | p |
| 0 | q |

Средняя арифметическая альтернативного признака

, т.к p+q=1.

Дисперсия альтернативного признака

, т.к1-р=q

Таким образом, дисперсия альтернативного признака равна произведению доли единиц, обладающих данным признаком, и доли единиц, не обладающих этим признаком.

Если значения 1 и 0 встречаются одинаково часто, т.е. p=q, дисперсия достигает своего максимума pq=0,25.

Дисперсия альтернативного признака используется в выборочных обследованиях, например, качества продукции.

## 2.1.3 Межгрупповая дисперсия. Правило сложения дисперсий

Дисперсия, в отличие от других характеристик вариации, является аддитивной величиной. То есть в совокупности, которая разделена на группы по факторному признаку х**,** дисперсия результативного признака y может быть разложена на дисперсию в каждой группе (внутригрупповую) и дисперсию между группами (межгрупповую). Тогда, наряду с изучением вариации признака по всей совокупности в целом, становится возможным изучение вариации в каждой группе, а также между этими группами.

Общая дисперсияизмеряет вариацию признака у по всей совокупности под влиянием всех факторов, вызвавших эту вариацию (отклонения). Она равна среднему квадрату отклонений отдельных значений признака у от общей средней и может быть вычислена как простая или взвешенная дисперсия.

Межгрупповая дисперсия характеризует вариацию результативного признака у, вызванную влиянием признака-фактора х, положенного в основу группировки. Она характеризует вариацию групповых средних и равна среднему квадрату отклонений групповых средних от общей средней :

,

где - средняя арифметическая i-той группы;

 - численность единиц в i-той группе (частота i-той группы);

 - общая средняя совокупности.

Внутригрупповая дисперсия отражает случайную вариацию, т.е. ту часть вариации, которая вызвана влиянием неучтенных факторов и не зависит от признака-фактора, положенного в основу группировки. Она характеризует вариацию индивидуальных значений относительно групповых средних, равна среднему квадрату отклонений отдельных значений признака у внутри группы от средней арифметической этой группы (групповой средней) и вычисляется как простая или взвешенная дисперсия для каждой группы:

 или ,

где - число единиц в группе.

На основании внутригрупповых дисперсий по каждой группе можно определить общую среднюю из внутригрупповых дисперсий:

.

Взаимосвязь между тремя дисперсиями получила название правила сложения дисперсий, согласно которому общая дисперсия равна сумме межгрупповой дисперсии и средней из внутригрупповых дисперсий:

**Пример**. При изучении влияния тарифного разряда (квалификации) рабочих на уровень производительности их труда получены следующие данные.

Таблица 5 - Распределение рабочих по среднечасовой выработке.

|  |  |  |  |
| --- | --- | --- | --- |
| №п/п | Рабочие 4-го разряда | №п/п | Рабочие 5-го разряда |
| Выработкарабочего, шт.,  |  |  | Выработкарабочего, шт.,  |  |  |
| 123456 | 799101213 | 7-10=-39-10=-11023 | 911049 | 1234 | 14141517 | 14-15=-1102 | 1104 |
|  | 60 | - | 24 | Σ | 60 | - | 6 |

В данном примере рабочие разделены на две группы по факторному признаку х- квалификации, которая характеризуется их разрядом. Результативный признак - выработка - варьируется как под его влиянием (межгрупповая вариация), так и за счет других случайных факторов (внутригрупповая вариация). Задача заключается в измерении этих вариаций с помощью трех дисперсий: общей, межгрупповой и внутригрупповой.

Определяем групповые и общую средние выработки, шт:

по первой группе шт.,

по второй группе шт.,

по двум группам шт.

Рассчитываем и заносим в таблицу и .

Рассчитываем внутригрупповые дисперсии:

по первой группе ,

по второй группе

Внутригрупповые дисперсии показывают вариации выработки в каждой группе, вызванные всеми возможными факторами (состояние оборудования, обеспеченность материалами и инструментами, возраст рабочих и т.д.), кроме различий в квалификации, т.к внутри группы все рабочие имеют одинаковый разряд.

Вычисляем среднюю из внутригрупповых дисперсий:

Средняя дисперсия отражает вариацию выработки, обусловленную всеми факторами, кроме квалификации, но в среднем по совокупности.

Межгрупповая дисперсия, характеризует вариацию среднегрупповых выработок, вызванную различием групп рабочих по квалификационному разряду:

Вычисляем общую дисперсию совокупности, которая отражает суммарное влияние всех возможных факторов на общую вариацию выработки изделий всеми рабочими:

Определяем общую дисперсию по правилу сложения дисперсий:

Очевидно, что чем выше доля межгрупповой дисперсии в общей дисперсии , тем сильнее влияние факторного признака (разряда) на результативный (выработку).

Эта доля характеризуется эмпирическим коэффициентом детерминации:

Эмпирический коэффициент детерминации показывает долю вариации результативного признака у под влиянием факторного признака х. Остальная часть общей вариации у вызвана изменением прочих факторов.

В примере эмпирический коэффициент детерминации равен:

 или 66,7%,

Это означает, что на 66,7% вариация производительности труда рабочих обусловлена различиями в квалификации, а на 33,3% - влиянием прочих факторов.

Эмпирическое корреляционное отношение показывает тесноту связи между группировочным и результативными признаками. Рассчитывается как корень квадратный из эмпирического коэффициента детерминации:

Эмпирическое корреляционное отношение , как и , может принимать значения от 0 до 1.

Если связь отсутствует, то =0. В этом случае =0, то есть групповые средние равны между собой и межгрупповой вариации нет. Значит группировочный признак - фактор не влияет на образование общей вариации.

Если связь функциональная, то =1. В этом случае дисперсия групповых средних равна общей дисперсии (), то есть внутригрупповой вариации нет. Это означает, что группировочный признак полностью определяет вариацию изучаемого результативного признака.

Чем ближе значение корреляционного отношения к единице, тем теснее, ближе к функциональной зависимости связь между признаками.

Для качественной оценки тесноты связи между признаками пользуются соотношениями Чэддока.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 0-0,2 | 0,2-0,3 | 0,3-0,5 | 0,5-0,7 | 0,7-0,9 | 0,9-0,99 | 1 |
| Сила связи | отсутствует | очень слабая | слабая | умеренная | заметная | тесная | весьма тесная | функцио-нальная |

В примере , что свидетельствует о тесной связи между производительностью труда рабочих и их квалификацией.


## 2.2 Относительные характеристики вариации

При сравнении вариации различных признаков или одного признака в различных совокупностях, используют относительные характеристики вариации - коэффициенты вариации.

Коэффициенты вариации рассчитываются как отношение абсолютных характеристик вариации (R,d,σ) к центру распределения и часто выражаются процентами. Линейный коэффициент вариации: . Квадратичный коэффициент вариации: . Коэффициент осциляции:

**Квадратичный коэффициент вариации используют как критерий однородности совокупности. Совокупность считается однородной, если**

Если центр распределения представлен медианой, то используют квартильный коэффициент вариации:


## 3. Теоретические кривые распределения

В вариационных рядах распределения существует определенная связь между изменением частот и значения варьирующего признака: частоты с ростом значения признака сначала увеличиваются, а затем после достижения какой-то максимальной величины в середине ряда уменьшаются. Значит, частоты в рядах изменяются закономерно в связи с изменением варьирующего признака. Такого рода закономерные изменения частот в вариационных рядах называются закономерностями распределения.

Анализ вариационных рядов предполагает выявление такой закономерности распределения, определение ее типа и построение теоретической кривой распределения, характеризующей данный тип распределения. Под кривой распределения понимают графическое изображение в виде непрерывной линии изменения частот в вариационном ряду, функционально связанного с изменением вариант. Эмпирической (фактической) кривой распределения является полигон. Под теоретическим распределением понимают вероятностное распределение частот в наблюдаемом вариационном ряду.

В практике статистического исследования встречаются распределения: нормальное, логарифмическое, биноминальное, Пуассона и др.

## 3.1 Нормальное распределение

При построении статистических моделей наиболее часто применяется нормальное распределение. Распределение непрерывной случайной величины х называют нормальным, если описывается следующей кривой:

где - ордината кривой нормального распределения (частости);

е=2,7182 - основание натурального логарифма;

π=3,1415 - постоянное число:

- нормированное отклонение.

Кривая нормального распределения симметрична относительно , поэтому величину называют центром распределения. На ее вид влияют значения и σ. Чем больше σ при неизменной , тем более плоской и растянутой вдоль оси абсцисс становится кривая, и наоборот.

Если σ остается неизменной, а изменяется, то кривые нормального распределения имеют одинаковую форму, но отличаются положением максимальной ординаты.

Особенности кривой нормального распределения (рис.2):

Кривая симметрична и имеет максимум в точке, где .

Кривая асимптотически приближается к оси абсцисс, продолжаясь в обе стороны до бесконечности.

Кривая имеет две точки перегиба при t = ±1, т.е. при таких значениях х, когда отклонение варианты от средней равно среднему квадратическому отклонению: .

При нормальном распределении 68,3% всех исследуемых частот находятся в пределах от до . В промежутке, ограниченном точками , находится 95,4%, а в промежутке , соответственно, 99,7% всех частот исследуемой совокупности (рис.1).

 y

 σ σ х

Рис.1. Кривая нормального распределения

## 3.2 Выравнивание эмпирического распределения по кривой нормального распределения

В анализе распределения большое значение имеет, насколько эмпирическое распределение признака соответствует нормальному. Для этого частоты фактического распределения нужно сравнить с теоретическими, которые характерны для нормального распределения. Значит, нужно по фактическим данным вычислить теоретические частоты кривой нормального распределения, являющиеся функцией нормированных отклонений (см. уравнение кривой ).

Иначе говоря, эмпирическую кривую распределения нужно выравнить кривой нормального распределения.

Порядок расчета теоретических частот кривой нормального распределения:

по эмпирическим данным рассчитывают среднюю арифметическую ряда и среднее квадратическое отклонение σ; находят нормированное отклонение t каждой варианты от средней арифметической; по таблице распределения функции определяют ее значения; вычисляют теоретические частоты по формуле:

,

где N - объем совокупности,

і - длина интервала;

строят и сравнивают графики эмпирические и теоретических частот (кривых распределения).

Сумма теоретических и эмпирических частот должна быть равной, но может не совпадать из-за округлений в расчетах.

## 3.3 Критерии согласия

Так как все предположения о характере того или иного распределения - это гипотезы, то они должны быть подвергнуты статистической проверке с помощью критериев согласия, которые дают возможность установить, когда расхождения между теоретическими и эмпирическими частотами следует признать несущественными, т.е. случайными, а когда - существенными (неслучайными). Таким образом, критерии согласия позволяют отвергнуть или подтвердить правильность выдвинутой при выравнивании ряда гипотезы о характере распределения в эмпирическом ряду.

Существует ряд критериев согласия. Чаще применяют критерии Пирсона, Романовского и Колмогорова.

**Критерий согласия Пирсона** - один из основных:

где k - число групп, на которые разбито эмпирическое распределение,

- наблюдаемая частота признака в i-й группе,

- теоретическая частота.

Для распределения составлены таблицы, где указано критическое значение критерия согласия для выбранного уровня значимости и степеней свободы df. (или )

Уровень значимости - вероятность ошибочного отклонения выдвинутой гипотезы, т.е. вероятность того, что будет отвергнута правильная гипотеза. В статистике пользуются тремя уровнями: α= 0,10, тогда Р=0,90 (в 10 случаях их 100 может быть отвергнута правильная гипотеза); α= 0,05, тогда Р=0,95; α= 0,01, тогда Р=0,99.

Число степеней свободы df определяется как число групп в ряду распределения минус число связей: df = k -z. Под числом связей понимается число показателей эмпирического ряда, использованных при вычислении теоретических частот, т.е. показателей, связывающих эмпирические и теоретические частоты.

Например, при выравнивании по кривой нормального распределения имеется три связи:

; ; .

Поэтому при выравнивании по кривой нормального распределения число степеней свободы определяется как df = k -3.

Для оценки существенности расчетное значение сравнивается с табличным .

При полном совпадении теоретического и эмпирического распределений , в противном случае >0. Если >, то при заданном уровне значимости и числе степеней свободы гипотезу о несущественности (случайности) расхождений отклоняем.

В случае, если , заключаем, что эмпирический ряд хорошо согласуется с гипотезой о предполагаемом распределении и с вероятностью Р= (1-α) можно утверждать, что расхождение между теоретическими и эмпирическими частотами случайно.

Критерий согласия Пирсона используется, если объем совокупности достаточно велик , при этом частота каждой группы должна быть не менее 5.

**Критерий Романовского с** основан на использовании критерия Пирсона, т.е. уже найденных значений , и числа степеней свободы df:

Он удобен при отсутствии таблиц для .

Если с<3, то расхождения распределений случайны, если же с>3, то не случайны и теоретическое распределение не может служить моделью для изучаемого эмпирического распределения.

**Критерий Колмогорова λ** основан на определении максимального расхождения между накопленными частотами и частостями эмпирических и теоретических распределений:

 или ,

где D и d - соответственно максимальная разность между накопленными частотами и накопленными частостями эмпирического и теоретического рядов распределений;

N - число единиц совокупности.

Рассчитав значение λ, по таблице Р (λ) определяют вероятность, с которой можно утверждать, что отклонения эмпирических частот от теоретических случайны. Вероятность Р (λ) может изменяться от 0 до 1. При Р (λ) =1 происходит полное совпадение частот, Р (λ) =0 - полное расхождение. Если λ принимает значения до 0,3, то Р (λ) =1.

Основное условие использования критерия Колмогорова - достаточно большое число наблюдений.

## 3.4 Характеристики неравномерности распределения

Симметричный вариационный ряд - это ряд, в котором частоты вариант, равностоящих от средней влево и вправо, равны между собой.

Необходимым, но недостаточным условием симметричности является равенство трех характеристик: средней арифметической, моды и медианы: = Ме=Mо

Этим соотношением пользуются для распознавания симметричности вариации.

Нормальное распределение, как отмечалось, характеризуется симметричностью. Поэтому сравнение фактического распределения с нормальным прежде всего констатирует отсутствие или наличие в нем асимметрии распределения. Асимметричные распределения встречаются чаще, чем симметричные.

Асимметричный вариационный ряд - это ряд, в котором частоты вариант, равностоящих от средней влево и вправо, не равны между собой и изменяются по-разному. Часто такой ряд называют скошенным

Различают правостороннюю и левостороннюю асимметрию (скошенность).

Ряд с правосторонней асимметрией имеет такой вид распределения частот

В рядах с правосторонней асимметрией >Ме>Mо, то есть наименьшим является значение моды, а наибольшим - средней.

Ряд с левосторонней асимметрией имеет такой вид распределения частот:

В рядах с левосторонней асимметрией < Ме< Mo, то есть наименьшим является значение средней, а наибольшим - моды.

Как видно из приведенных рисунков, асимметрию легко определить визуально по виду полигона или гистограммы распределения. При левосторонней асимметрии относительно центра распределения наблюдается длинная левая ветвь кривой распределения, тогда как при правосторонней асимметрии - правая ветвь этой кривой.

В качестве показателя асимметрии применяется коэффициент асимметрии Пирсона:

.

Если Ка >0, скошенность правосторонняя, если Ка<0, скошенность левосторонняя; если Ка=0, вариационный ряд симметричен.

Кроме симметричности расположения кривой относительно ординаты средней арифметической, сравнение фактического распределения с нормальным производится и на эксцесс. Под эксцессом распределения понимается высоковершинность или, наоборот, низковершинность фактической кривой распределения по сравнению с нормальным распределением:

Высоковершинность означает положительный эксцесс и характеризует скопление частот в середине. Низковершинность означает отрицательный эксцесс и большую разбросанность членов ряда.