**Лабораторная работа №2  
по курсу «Компьютерный анализ статистических данных»  
на тему «Гистограмма распределения»**

## Цель работы

Изучить правила построения гистограммы и кумуляты распределения случайных величин, сравнивать гистограмму с нормальным и равномерным законами распределения. Научиться построению гистограмм распределения средствами MS Excel.

## Теоретические сведения

Гистограмма распределения показывает, насколько часто в выборке встречаются те или иные значения X (аналог плотности распределения). Каждый столбец гистограммы показывает частоту попадания значения выборки в интервал значений – чем выше столбец, тем вероятнее соответствующие значения показателя.

Для построения гистограммы необходимо:

1. Разбить диапазон значений показателя на интервалы.
2. Посчитать число значений, попавших в каждый интервал – *частоты*.
3. Вычислить высоту столбцов гистограммы в соответствии с правилом нормировки: сумма площадей столбцов гистограммы равна 1 (без учета зазоров между столбцами).
4. Построить график.

Рассмотрим каждый шаг подробнее.

### Выделение интервалов

Весь диапазон значений показателя необходимо разбить на части – ***интервалы***, или «***карманы***». Не допускается наличие «дырок» между интервалами и перекрытие интервалов. Чаще всего интервалы имеют одинаковую ширину.

Интервал задается своей нижней и верхней границей: *aj* и *bj*, *j* = 1,2,3... – номер интервала.

Проще всего выделить *интервалы* *с «круглыми» границами*, например, шириной 10:

1. [20;30] *a*1 = 20 *b*1 = 30
2. (30;40] *a*2 = 30 *b*2 = 40
3. (40;50] *a*3 = 40 *b*3 = 50
4. (50;60] *a*4 = 50 *b*4 = 60
5. (60;70] *a*5 = 60 *b*5 = 70

Обратите внимание, что конец одного интервала – это начало следующего интервала, т.е. *bj* = *aj*+1: *b*1 = *a*2, *b*2 = *a*3 и т.д.

Но в этом случае не понятно, какой именно ширины брать интервалы: 10, 50, 100? Если взять слишком маленькие, то в каждый будет попадать мало значений (1-2 или вообще 0), а если слишком большие, то наоборот – слишком много.

Иногда разбиение выполняют в соответствии *с общепринятой традицией*. Например, во всем мире принято распределять население по возрастным группам по 5 лет, как показано в таблице ниже.

**Распределение населения РФ по возрастным группам**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Тыс. человек | | |
| 2001 | 2006 | 2007 |
| **Все  население** | **146304** | **142754** | **142221** |
| в том числе    в возрасте, лет**:** |  |  |  |
| 0-4 | 6367 | 7037 | 7223 |
| 5-9 | 7762 | 6418 | 6376 |
| 10-14 | 11789 | 7790 | 7283 |
| 15-19 | 12322 | 11825 | 11088 |
| 20-24 | 11106 | 12405 | 12671 |
| 25-29 | 10451 | 11049 | 11165 |
| 30-34 | 9620 | 10295 | 10442 |
| 35-39 | 11333 | 9417 | 9459 |
| 40-44 | 12651 | 10949 | 10368 |
| 45-49 | 11434 | 12054 | 12067 |
| 50-54 | 9409 | 10645 | 10804 |
| 55-59 | 4995 | 8590 | 8985 |
| 60-64 | 8906 | 4407 | 4336 |
| 65-69 | 5903 | 7609 | 7458 |
| 70 и более | 12256 | 12264 | 12496 |

*Источник: gks.ru*

Третий вариант – выяснить, какое *количество* *интервалов* является оптимальным исходя из мощности выборки (чем больше наблюдений, тем больше интервалов). Для этого используется формула Стерджесса:



Здесь *m* – количество интервалов (не путать со средним значением *mx*).

Примеры

*n* = 10 *m* = 1 + 3,322 lg 10 ≈ 4,3 *m* = 4

*n* = 20 *m* = 1 + 3,322 lg 20 ≈ 5,3 *m* = 5

*n* = 30 *m* = 1 + 3,322 lg 30 ≈ 5,9 *m* = 6

*n* = 40 *m* = 1 + 3,322 lg 40 ≈ 6,3 *m* = 6

*n* = 50 *m* = 1 + 3,322 lg 50 ≈ 6,6 *m* = 7

*n* = 100 *m* = 1 + 3,322 lg 100 ≈ 7,6 *m* = 8

*n* = 1000 *m* = 1 + 3,322 lg 1000 ≈ 10,97 *m* = 11

*n* = 10000 *m* = 1 + 3,322 lg 10000 ≈ 14,8 *m* = 15

Затем весь диапазон значений от минимального до максимального разбивается *на* *равные* *части*. Ширина интервала *w* определяется по формуле:



Первый интервал начинается в *X*min, а последний (с номером *m*) заканчивается в *X*max.

*a*1 = *X*min *b*1 = *a*1 + *w*

*a*2 = b1 *b*2 = *a*2 + *w*

…

*am* = *bm*–1 *bm* = *am* + *w* = *X*max

Пример

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *i* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *X* | 9 | 10 | 10 | 12 | 14 | 15 | 18 | 19 | 21 | 22 | 28 | 37 | 43 | 44 | 47 |

*n* = 15

*m* = 1 + 3,322 lg 15 = 4,9 = 5

*X*min = 9

*X*max = 47

*w* = (47 – 9) / 5 = 38 / 5 = 7,6

*a*1 = *X*min = 9 *b*1 = 9 + 7,6 = 16,6

*a*2 = 16,6 *b*2 = 16,6 + 7,6 = 24,2

*a*3 = 24,2 *b*3 = 24,2 + 7,6 = 31,8

*a*4 = 31,8 *b*4 = 31,8 + 7,6 = 39,4

*a*5 = 39,4 *b*5 = 39,4 + 7,6 = 47 = *X*max

Интервалы в виде таблицы:

|  |  |  |
| --- | --- | --- |
| *j* | *a* | *b* |
| 1 | 9 | 16,6 |
| 2 | 16,6 | 24,2 |
| 3 | 24,2 | 31,8 |
| 4 | 31,8 | 39,4 |
| 5 | 39,4 | 47 |

### Подсчет попаданий в интервалы

**Частота (частость)** *nj* попадания в каждый *j*-тый интервал считается явно: сколько из всех значений находятся внутри интервала, т.е.  и .

Сумма всех частот должна быть равна мощности выборки, иначе какие-то значения «потерялись»:



В примере с населением частота попадания в интервал 20-24 года в 2007г. составила 12671 тыс. чел.

Пример (продолжение)

|  |  |  |  |
| --- | --- | --- | --- |
| *j* | *a* | *b* | *n* |
| 1 | 9 | 16,6 | 6 |
| 2 | 16,6 | 24,2 | 4 |
| 3 | 24,2 | 31,8 | 1 |
| 4 | 31,8 | 39,4 | 1 |
| 5 | 39,4 | 47 | 3 |

Сумма *nj*: 6 + 4 + 1 + 1 + 3 = 15 = *n*

**Высота столбца** гистограммы вычисляется через частоту и ширину интервалов:



При этом выполняется *правило нормировки* – площадь графика плотности распределения равна 1.

Доказательство

Площадь каждого столбца (прямоугольника) – это высота умножить на ширину:



Сумма площадей:



Пример (продолжение)

*h*1 = 6 / (15 ∙ 7,6) = 6 / 114 ≈ 0,05263

*h*2 = 4 / (15 ∙ 7,6) = 4 / 114 ≈ 0,03509

*h*3 = 1 / (15 ∙ 7,6) = 1 / 114 ≈ 0,00877

*h*4 = *h*3 ≈ 0,00877

*h*5 = 3 / (15 ∙ 7,6) = 3 / 114 ≈ 0,02632

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *j* | *a* | *b* | *n* | *h* |
| 1 | 9 | 16,6 | 6 | 0,05263 |
| 2 | 16,6 | 24,2 | 4 | 0,03509 |
| 3 | 24,2 | 31,8 | 1 | 0,00877 |
| 4 | 31,8 | 39,4 | 1 | 0,00877 |
| 5 | 39,4 | 47 | 3 | 0,02632 |

### Анализ графика гистограммы

Для визуального анализа гистограмму наносят на график и сравнивают с известными законами распределения.

Чаще всего гистограмма близка к **нормальному закону распределения**. Большинство значений сосредоточены в центре, к краям высоты столбцов плавно снижается. Пример такой гистограммы в сравнении с теоретическим нормальным законом (красная линия):

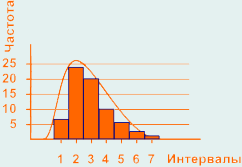
Другой возможный вариант – **равномерное распределение**, когда все значения примерно поровну распределены по столбцам гистограммы, нет заметных максимумов и минимумов:

В идеале, высоты столбцов должны быть почти одинаковыми, но в реальности так получается редко.

Существуют и другие варианты гистограммы распределения:

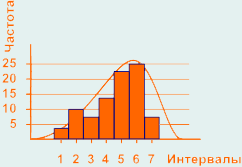
**Положительная асимметрия**

Гистограмма смещена влево, длинный правый «хвост».

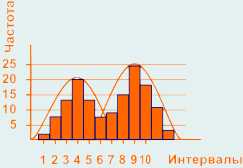


**Отрицательная асимметрия**

Гистограмма смещена вправо, длинный левый «хвост».



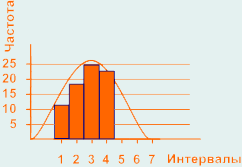
**Бимодальность**



Гистограмма отображает два совмещенных процесса. Такая ситуация может произойти если в исследуемой совокупности смешались два разных по природе типа объектов.

Например, собрана общая статистика спортивных результатов для мужчин и женщин, хотя они сильно различаются.

**Гистограмма усечена**



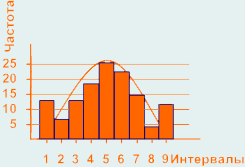
Распределение не является нормальным т.к. нет постепенного снижения частоты результатов измерений от центра к границам. Такой вид гистограммы возникает, если чересчур малые и большие наблюдения невозможны или потеряны.

**Гистограмма не имеет центра**



Достаточно редкая ситуация, когда столбцы в центре диаграммы близки к 0. Может быть случаем бимодальности. Такая гистограмма имеет отрицательный эксцесс.

**«Тяжелые хвосты»**



По краям гистограммы наблюдаются высокие столбцы. Означает сильное «расслоение» показателя, когда много как очень больших, так и очень маленьких значений. Коэффициент эксцесса отрицательный.

Пример (продолжение)

Для нашего примера гистограмма имеет вид:

Она не соответствует ни нормальному, ни равномерному закону распределения: столбцы по краям гистограммы выше, чем посередине. Присутствуют «тяжелые хвосты», гистограмма смещена влево (положительная асимметрия), коэффициент эксцесса отрицательный.

### Кумулятивная кривая

**Кумулята** или **кумулятивная кривая,** в отличие от гистограммы строится по накопленным частотам.

*Накопленная частота* – это сумма частот обычной гистограммы, от первого интервала до текущего:







...



Другими словами, это количество наблюдений, меньше верхней границы интервала (<*bj*). Последняя накопленная частота всегда равна *n*, т.к. это число попаданий во все интервалы, т.е. все наблюдения.

Высота столбца кумуляты рассчитывается как накопленная частота, деленная на общее количество наблюдений:





На графике на горизонтальной оси также отмечают интервалы значений, а на вертикальной оси – накопленные частоты. Кумулята является аналогом закона распределения, и поэтому возрастает от 0 до 1.

Визуально кумуляту труднее анализировать, но с вертикальной оси можно снимать значения вероятностей: высота столбца кумуляты равна вероятности того, что случайная величина меньше верхней границы соответствующего интервала.



Пример (продолжение)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *j* | *a* | *b* | *n* | *h* | *k* | *H* |
| 1 | 9 | 16,6 | 6 | 0,05263 | 6 | 0,400 |
| 2 | 16,6 | 24,2 | 4 | 0,03509 | 10 | 0,667 |
| 3 | 24,2 | 31,8 | 1 | 0,00877 | 11 | 0,733 |
| 4 | 31,8 | 39,4 | 1 | 0,00877 | 12 | 0,800 |
| 5 | 39,4 | 47 | 3 | 0,02632 | 15 | 1,000 |

График кумуляты:

Таким образом, с вероятностью 0,4 (40%) значения исследуемого показателя будут меньше 16,6; с вероятностью 0,667 (66,7%) – меньше 24,2 и т.д.

## Задание

Имеется статистическая выборка с наблюдениями показателя *X –* суммы ежемесячного заработка в определенной профессии, тыс. д.е. Необходимо построить гистограмму распределения:

1. Определить мощность выборки, минимальное, максимальное, среднее значения и стандартное отклонение.
2. Определить число интервалов по формуле Стерджесса и их ширину.
3. Определить границы и середины интервалов («карманов»).
4. Вычислить частоты попаданий в интервалы двумя способами: через «Анализ данных» и через формулы.
5. Построить график гистограммы.
6. Построить график кумуляты.
7. Сравнить полученную гистограмму с нормальным законом распределения (визуально).
8. Сравнить полученную гистограмму с равномерным законом распределения (визуально).

Сделать выводы о предположительном законе распределения (нормальный, равномерный, другой). Если это возможно, определить, какие суммы заработка а) наиболее вероятны; б) наименее вероятны. Какова вероятность иметь ежемесячный заработок менее 10 тыс.д.е. в данной профессии?

## Пример

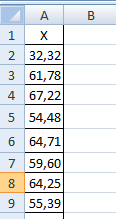
Все расчеты показаны на примере выборки:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 32,32 | 43,69 | 55,62 | 33,28 | 59,92 |
| 61,78 | 39,15 | 39,94 | 51,83 | 55,95 |
| 67,22 | 52,41 | 40,67 | 55,92 | 42,29 |
| 54,48 | 40,87 | 57,46 | 49,89 | 59,24 |
| 64,71 | 45,79 | 30,92 | 34,2 | 56,32 |
| 59,6 | 51,18 | 46,25 | 37,62 | 63,64 |
| 64,25 | 46,88 | 43,76 | 48,74 | 48,74 |
| 55,39 | 56,84 | 40,61 | 51,05 | 56,89 |
| 50,88 | 59,99 | 51,16 | 43,99 | 61,4 |
| 60,91 | 56,38 | 63,22 | 46,57 | 42,99 |
| 49,05 | 58,19 | 44,96 | 53,84 | 37,59 |
| 53,29 | 56,08 | 68,03 | 58,3 | 42,58 |
| 57,67 | 54,15 | 50,71 | 56,01 | 46,73 |
| 31,14 | 59,29 | 57,91 | 44,76 | 28,71 |
| 50,4 | 44,7 | 41,5 | 46,79 | 23,02 |
| 53,17 | 44,9 | 54,87 | 65,09 | 26,01 |
| 55,89 | 62,02 | 59,35 | 36,57 | 32,53 |

## Подготовка исходных данных

Работа выполняется в том же файле, что и первая, но на новом листе. Переименуйте «Лист 2» в «Л.р.2».

Скопируйте данные для своего варианта на лист «Л.р.2», начиная со второй строки и столбца A. Укажите название столбца (X).



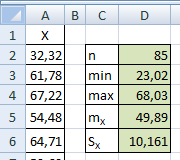
Нумеровать значения в данной работе необязательно.

## Указания к выполнению работы

В описании работы для наглядности все ячейки с расчетами закрашены зеленым. Вам это делать необязательно.

1. Сначала необходимо рассчитать основные числовые характеристики: мощность выборки, минимальное и максимальное значения, среднее и стандартное отклонение. Они потребуются ходе выполнения работы и для выводов.

Числовые характеристики рассчитываются по формулам из лабораторной работы №1 (любым способом).



В примере выборка содержит 85 наблюдений, от 23,02 до 68,03. Средний доход составляет 49,89, стандартное отклонение – 10,16.

Отчет

Добавьте в отчет скриншот с основными числовыми характеристиками, опишите их значения в выводе.

1. Начнем построение гистограммы – вычислим количество интервалов и их ширину.

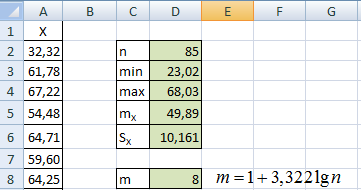
Число интервалов определим по формуле Стерджесса:



Для удобства, формулу можно скопировать на лист Excel.

Полученное число необходимо округлить до большего значения.

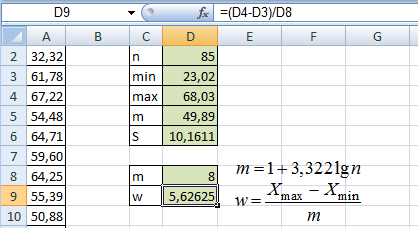
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Функции:   |  |  | | --- | --- | | ОКРУГЛВВЕРХ(*число*; *знаков\_после\_запятой*) | Округлить число в сторону большего до указанного знаков после запятой (0 – до целых). | | LOG10(*число*) | Десятичный логарифм числа (lg). | |



Чтобы определить ширину интервала, воспользуемся формулой:



Ширину интервала округлять **не** нужно.



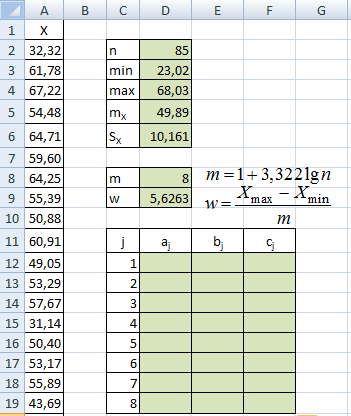
Таким образом, необходимо исходные данные распределить по 8 интервалам шириной 5,626 тыс.д.е.

Отчет

Добавьте в отчет скриншот и формулы для расчета количества и ширины интервалов.

1. Определим границы интервалов («карманов») и подготовим таблицу для будущих расчетов. Число строк в таблице равно числу интервалов, столбцы мы будем добавлять по мере необходимости (всего их будет 9, пока подготовим 3).

|  |  |
| --- | --- |
|  | Таблицу необходимо просто расчертить, **не надо** вставлять ее через команду на Ленте. |



*aj* – нижняя граница;

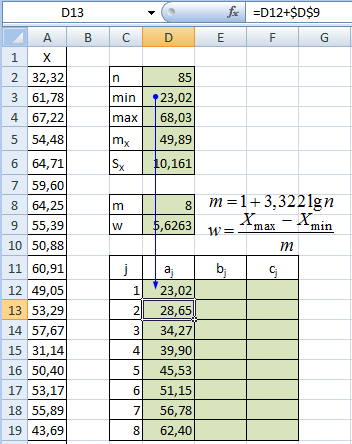
*bj* – верхняя граница;

*cj* – середина интервала.

Нижняя граница первого интервала – это минимальное значение выборки (в примере 23,02). Нижняя граница каждого следующего интервала – это предыдущий интервал плюс ширина интервала:

.

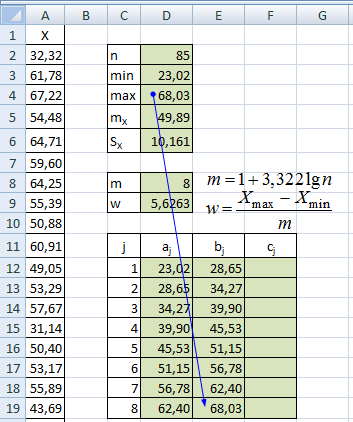
|  |  |
| --- | --- |
|  | Не забудьте зафиксировать ячейку с **w** в формуле, прежде чем растянуть ее! |



Верхняя граница интервала совпадает с нижней границей следующего интервала и рассчитывается по такой же формуле.

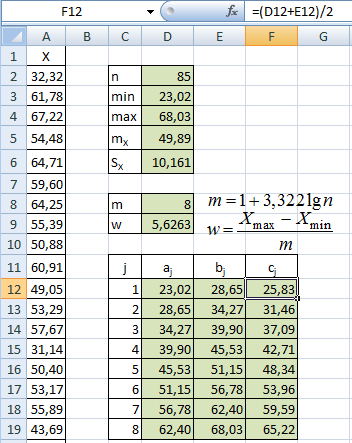
.

Верхняя граница последнего интервала должна быть в точности равна максимальному значению выборки.



Середина интервала рассчитывается как среднее между его границами.





Таким образом, исходную выборку мы разбили на 8 числовых интервалов:

1. [23,02;28,65]
2. (28,65;34,27]
3. (34,27;39,9]
4. (39,9;45,53]
5. (45,53;51,15]
6. (51,15;56,78]
7. (56,78;62,4]
8. (62,4;68,03]

Квадратная скобка означает, что число на границе входит в интервал, круглая – что не входит. Например, 28,65 входит в первый интервал, но не входит во второй. На практике иногда делают наоборот, в зависимости от удобства.

Обратите внимание, в первый интервал входят обе границы.

1. Далее необходимо определить частоты попадания в интервал, т.е. сколько наблюдений из исходной выборки попадает в каждый интервал.

|  |  |
| --- | --- |
|  | В Excel это можно сделать двумя способами: через «Анализ данных» и через формулы. |

Способ 1.

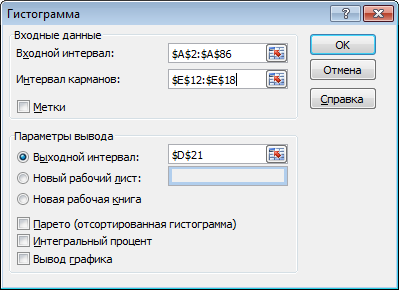
Убедитесь, что на вкладке «Данные» есть кнопка «Анализ данных» (как ее включить см. в л.р.1).

Нажмите «Анализ данных» – «Гистограмма» – «ОК».

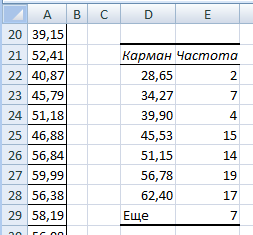
В поле «Входной интервал» укажите все значения X.

В поле «Интервал карманов» – верхние границы интервалов, кроме последней.

В «Параметрах вывода» выберите «Выходной интервал» и укажите любую ячейку под таблицей. В качестве «карманов» указаны верхние границы интервалов, кроме последнего.



В результате получим таблицу вида:

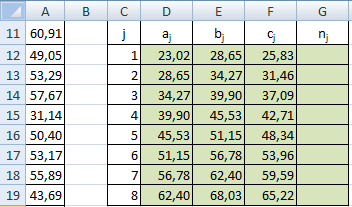


Карман «Еще» соответствует последнему интервалу и считает все значения, не вошедшие в остальные интервалы.

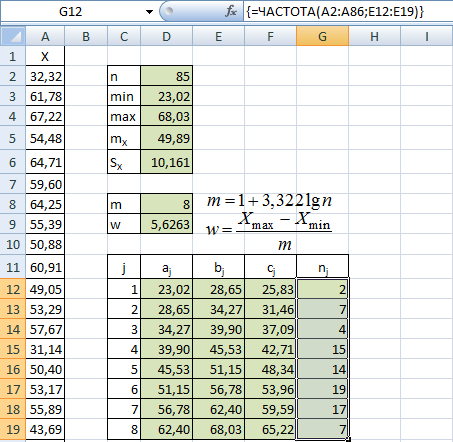
Способ 2.

Способ более трудоемкий, но позволяет автоматически пересчитывать гистограмму при изменении или добавлении новых значений.

Значения по способу 2 мы внесем в следующий столбец нашей таблицы, обозначив его *nj*.

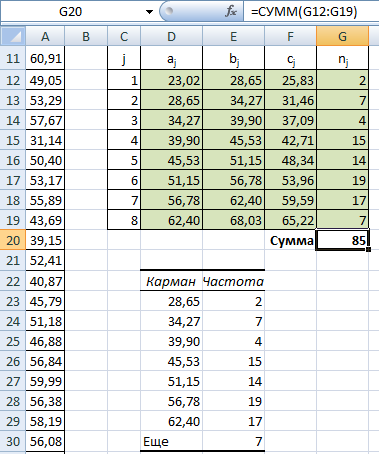


|  |  |
| --- | --- |
|  | Воспользуемся формулой  ЧАСТОТА(*исходные\_данные*;*интервалы*)  *исходные\_данные* – это все значения X  *интервалы* – верхние границы интервалов  Внимание!  Это одна из специальных формул Excel – формула массива. Результат расчета по такой формуле занимает несколько ячеек. Она **не** растягивается, и **не** вводится отдельно в каждую ячейку, а вводится следующим образом.   1. Выделите весь диапазон ячеек, в которых будут значения *nj* (в примере G12:G19). 2. Нажмите на клавиатуре клавишу F2. Включится режим ввода формулы, но выделены будут по-прежнему все ячейки. 3. Введите формулу для расчета. В примере =ЧАСТОТА(A2:A86;E12:E19). **Не** нажимайте Enter. 4. Нажмите сочетание Ctrl+Shift+Enter.   В результате формула заполнит все ячейки, в строке формул она отображается в фигурных скобках {=ЧАСТОТА(A2:A86;E12:E19)}  Чтобы отредактировать такую формулу нужно повторить ту же последовательность действий: выделить все ячейки → F2 → ввод формулы → Ctrl+Shift+Enter.  При попытке изменить отдельную ячейку внутри массива, Excel выдаст ошибку. |



Проверьте себя – сумма *nj* должна быть равна *n*. Добавьте под таблицей строку «Сумма» и вычислите в ней сумму *nj*.

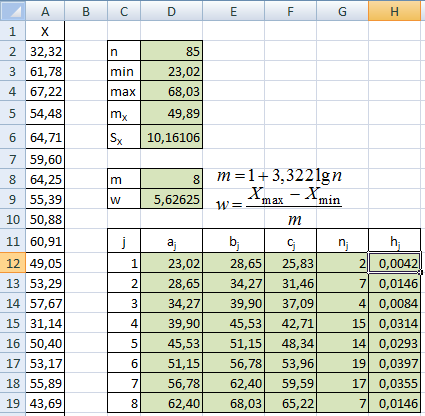
Результат для обоих способов должен совпадать.



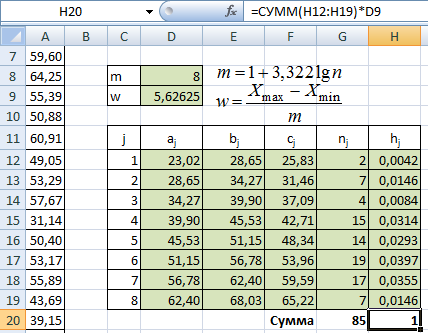
1. Построим график гистограммы. Для этого потребуются некоторые дополнительные расчеты, т.к. частоты – это еще не гистограмма. Для нее должно выполняться правило нормировки – сумма площадей всех столбцов гистограммы должна быть равна 1. Исходя из этого правила, высота столбца гистограммы вычисляется по формуле:



Добавим соответствующие расчеты.



Проверьте себя! Сумма *hj*, умноженная на *w*, должна быть равна 1.



Построим график гистограммы.

|  |  |
| --- | --- |
|  | Выделите значения *hj* вместе с заголовком и на вкладке «Вставка» выберите «Диаграмма» – «Гистограмма» – «Гистограмма с группировкой» (самая первая).  Удалите заголовок диаграммы.  Чтобы под столбцами отображались не номера интервалов, а их границы, нужно кликнуть по гистограмме правой кнопкой – «Выбрать данные...», нажать кнопку «Изменить» справа, в разделе «Подписи горизонтальной оси» и выделить значения *aj* и *bj* (без заголовков). |

В результате получим график вида:

Отчет

Добавьте в отчет скриншот с расчетом частот и высоты столбцов, а также график гистограммы.

1. Построим кумулятивную кривую по сумме накопленных частот. Для кумуляты также отдельно рассчитываются накопленные частоты (*ki*) и высоты столбцов (*Hj*).

Частоты можно вычислить следующим образом.

Частота первого интервала одинакова для гистограммы и кумуляты:



Остальные частоты рассчитываются как предыдущая частота кумуляты плюс текущая частота гистограммы:



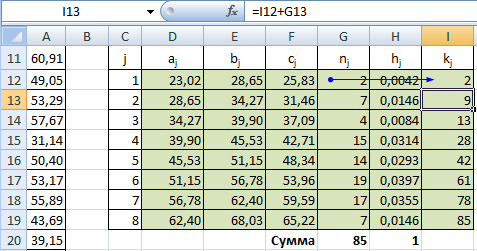
Например, для 2 и 3 строки получится:





и т.д.

Добавим в таблицу столбец *kj*:

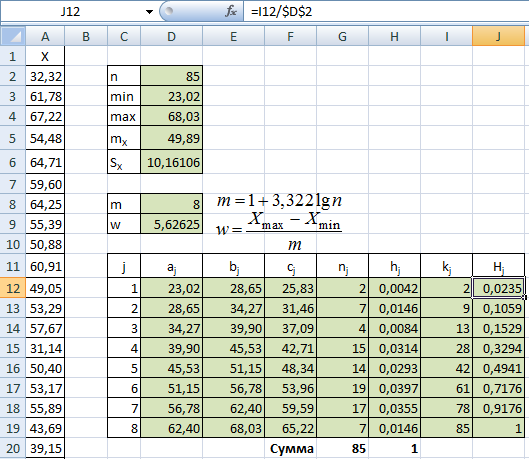


Обратите внимание, в последнем интервале должна получиться мощность выборки *n*.

Теперь можно вычислить высоты столбцов кумуляты:



Для кумуляты правило нормировки не действует, поэтому частоты делятся просто на *n*. В последней строке должна получиться 1.



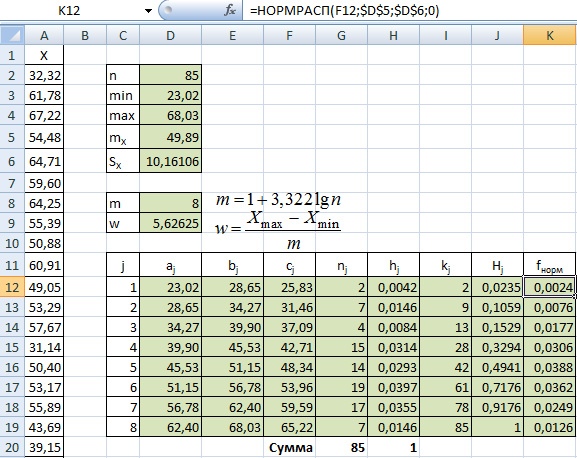
По значениям кумуляты можно сказать, например, что только в 2,35% случаев сумма ежемесячного заработка в данной профессии составляет менее 28,65 тыс.д.е. А почти в половине случаев (49,41%) сумма заработка составляет до 51,15 тыс.д.е.

График кумуляты строится аналогично обычной гистограмме. Можно скопировать уже созданную диаграмму и изменить исходные данные.

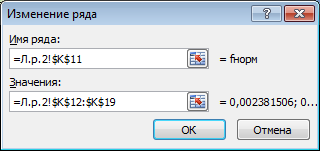
1. Сравним полученную в п.5 гистограмму с нормальным законом распределения, который очень часто встречается на практике.

Добавим в таблицу еще один столбец для расчета плотности нормального распределения.

|  |  |
| --- | --- |
|  | Воспользуемся формулой НОРМРАСП(*x*;*mx*;*Sx*;*тип*)  *x* – для какого значения рассчитываем (середины интервалов *cj*)  *mx*, *Sx* – среднее значение и стандартное отклонение  тип = 0, показывает, что нужно рассчитать именно плотность, а не интегральный закон (кумуляту) |

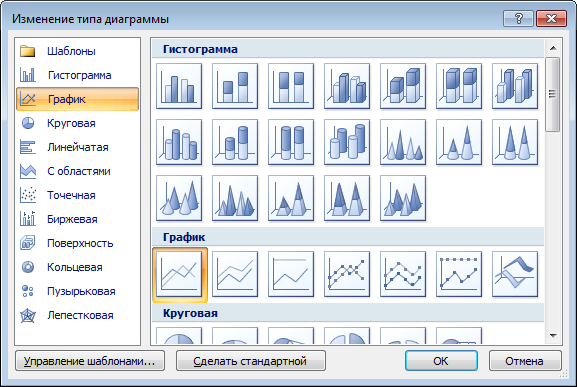


Добавим его на график с гистограммой. Кликните по ней правой кнопкой – «Выбрать данные», нажмите «Добавить» и укажите столбец с fнорм.



В результате диаграмма примет вид:

Но плотность распределения принято изображать не столбцами, а гладкой линией. Для этого выделите новые столбцы (красные), кликните по любому из них правой кнопкой – «Изменить тип диаграммы для ряда». Выберите тип – «График», нажмите «ОК». Столбцы превратятся в линию.



Еще раз кликните по полученной линии правой кнопкой – «Формат ряда данных» – «Тип линии», поставить галочку «сглаженная линия».В результате получим:

**Вывод**: В целом, гистограмма распределения не соответствует нормальному закону: она не симметрична (справа столбцы высокие, слева низкие, мода не в центре распределения). Имеются «провалы» в 3 и 5 интервалах, а 2 и 7 столбцы, наоборот, значительно выше нормальной плотности распределения.

Примечание

Более точно на этот вопрос позволяют ответить специальные критерии, например, критерий Пирсона, Колмогорова. Их расчет выходит за рамки данной лабораторной, но они тоже строятся на основе гистограммы.

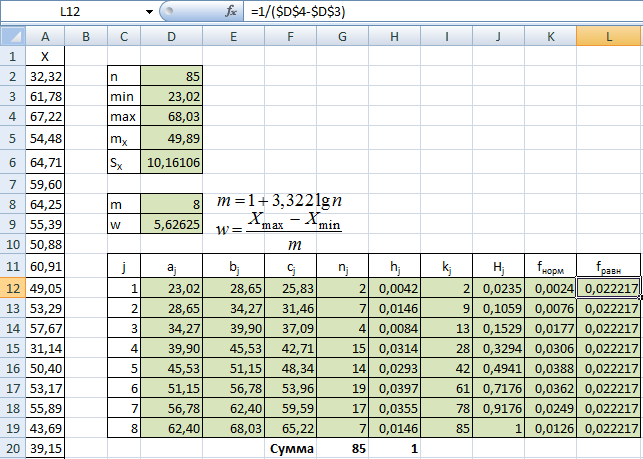
Отчет

Добавьте в вывод сравнение полученной гистограммы с нормальным законом распределения. График пока вставлять не нужно, на него необходимо добавить еще одну линию.

1. Сравним гистограмму с еще одним распространенным законом распределения – равномерным. У равномерного закона отсутствует мода, нет более или менее вероятных интервалов – все допустимые значения равновероятны.

Значения плотности равномерного распределения во всех точках одинаковы и рассчитываются по формуле:





Добавим равномерный закон на график гистограммы аналогично нормальному.

**Вывод**: Гистограмма распределения далека от равномерного закона. Заметно, что высоты столбцов сильно отличаются друг от друга. В правой части гистограммы столбцы выше, т.е. более высокий заработок вероятнее, чем низкий.

Отчет

Добавьте в отчет гистограмму в сравнении с нормальным и равномерным законом распределения. В выводе напишите, можно ли считать закон распределения равномерным.

Напишите общий вывод – охарактеризуйте гистограмму распределения в целом. Дайте ответы на вопросы из задания.

**Общий вывод:**

В рассматриваемой профессии наиболее вероятна сумма ежемесячного заработка от 50 до 60 тыс.д.е. (округленно). Маловероятна сумма заработка более 70 и 20 тыс.д.е.

Закон распределения нельзя считать ни нормальным, ни равномерным. Он асимметричный, смещен вправо, с модой в диапазоне от 51,15 до 56,78, что немного больше среднего.

Вероятность получения заработка менее 10 тыс. д.е. в данной профессии близка к 0.