

Обработка статистических данных средствами MS Excel. Построение гистограммы.

Исходным понятием статистики является понятие **генеральная совокупность**, объединяющее обычно какое-либо множество испытуемых (учащихся) по одному или нескольким интересующим признакам. Главное требование к выделению изучаемой совокупности — это ее качественная однородность, например, по уровню знаний, росту, весу и другим признакам.

Применение большинства статистических методов основано на идее использования небольшой случайной совокупности испытуемых из общего числа тех, на которых можно было бы распространить (генерализовать) выводы, полученные в результате изучения совокупности. Эта небольшая совокупность в статистике называется **выборочной совокупностью** (или короче — **выборкой**).

Любая статистическая работа предполагает выбор генеральной совокупности. Генеральная совокупность – совокупность однородных статистических данных, отличающихся друг от друга, но при этом имеющих сходные свойства. Но, с генеральной совокупностью на практике работать невозможно из-за огромного числа объектов, входящих в нее. Поэтому на практике формируют выборку. Выборка – часть генеральной совокупности. На выборке измеряют переменные.

Переменная – это то, что можно измерять, контролировать или чем можно манипулировать в исследованиях. Иными словами, переменная – это то, что варьируется, изменяется, а не остается постоянным. Например, измеряя давление или содержание лейкоцитов в крови можно получить различные значения у различных пациентов или значения для одного пациента в разное время суток.

Можно привести примеры разных переменных. И, понятно, что все они будут отличаться своими характеристиками.

Так как значения переменных не постоянны, нужно научиться описывать их изменчивость.

Для этого придуманы описательные статистики: минимум, максимум, среднее, дисперсия, стандартное отклонение, медиана, квартили, мода и т.д. Они дают общие представления о значениях, которые принимает переменная.

Минимум и максимум – это минимальное и максимальное значения переменной.

Среднее – это сумма всех значений переменной, деленное на количество переменных.

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

Дисперсия и стандартное отклонение - наиболее часто используемые меры изменчивости переменной.

Дисперсия рассчитывается по формуле:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Стандартное отклонение равно квадратному корню из дисперсии. Формально имеем:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Чем выше стандартное отклонение и дисперсия, тем сильнее разбросаны значения переменной относительно среднего

Медиана- это величина, относительно которой ряд распределения делится на две равные части: в обе стороны от медианы располагается одинаковое число вариант. При наличии небольшого числа вариант медиана определяется довольно просто. Для этого все данные нумеруют, и при нечетном числе вариант центральная варианта и будет его медианой. При четном числе определяется как среднее значение двух величин, находящихся в середине ряда. Для величин, по которым построена гистограмма, медиану можно определить следующим способом. Необходимо найти класс, в котором содержится медиана. Для этого необходимо складывать частоты встречаемости по классам до тех пор, пока сумма частот не превзойдет половину всех членов ряда. Данный класс называется медианным. Тогда медиану можно найти по формуле:

$$Me = x_n + \lambda \left(\frac{\frac{n}{2} - \sum f_i}{f_{Me}} \right)$$

где x_n - нижняя граница интервала, содержащего медиану, $\sum f_i$ - сумма накопленных частот, стоящая перед медианным классом, λ - величина классового интервала, f_{Me} – частота медианного класса, n - общее число наблюдений.

Мода- это величина, наиболее часто встречающаяся в данной совокупности. Класс с наибольшей частотой называется модальным. Моду можно найти по формуле:

$$Mo = x_n + \lambda \left(\frac{f_2 - f_1}{2f_2 - f_1 + f_3} \right)$$

где x_n - нижняя граница модального класса, f_2, f_1 - частота класса, предшествующего модальному, f_3 - частота класса, следующего за модальным, λ - ширина классового интервала.

Коэффициент асимметрии характеризует симметричность графика распределения относительно среднего значения. Он принимает значения от -1 до 1.

Коэффициент асимметрии определяется по формуле:

$$As = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^3}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^{\frac{3}{2}}}$$

Эксцесс, в свою очередь, характеризует узковершинность или плосковершинность распределения. Эксцесс определяется по формуле:

$$\Theta = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^4}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^2} - 3$$

Далее необходимо характеристики, полученные на выборке, перенести на всю генеральную совокупность. Для этого вводят понятие доверительного интервала.

Доверительный интервал- это интервал, в который попадает среднее значение генеральной совокупности с той или иной вероятностью.

$$\bar{x} - t \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t \frac{\sigma}{\sqrt{n}}$$

В качестве наглядного представления выборочной совокупности можно рассмотреть гистограмму. Гистограмма- это фигура, состоящая из прямоугольников, ширина которых- ширина класса, а высота- функция плотности вероятности.

Рассмотрим алгоритм построения гистограммы на конкретной задаче.

Дан рост группы детей. Согласно полученным данным построить гистограмму.


70 69 72 73 71 66 73 67 68 73
 71 69 67 74 71 70 70 67 71 69
 70 70 70 71 74 74 71 69 72 71

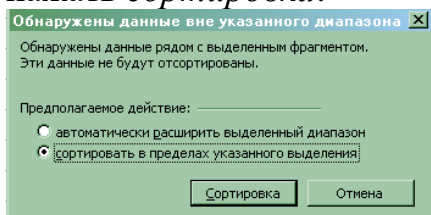
Ввод данных

1. В ячейку A1 ввести заголовок *РОСТ*
2. В ячейки A2-A31 ввести данные задачи.

Построение вариационного ряда

1. В ячейку B1 ввести заголовок *вариационный ряд*
2. Скопировать данные ячеек A2-A31 в ячейки B2-B31. Для этого выделить ячейки A2-A31, щелкнуть правой кнопкой мыши по выделенным ячейкам, выбрать пункт меню *копировать*, встать на ячейку B2, щелкнуть правой кнопкой мыши, выбрать пункт меню *вставить*.

3. Выделить ячейки В2-В31. На панели инструментов выбрать кнопку  сортировать по возрастанию, в появившемся окне выбрать предполагаемое действие *сортировать в пределах указанного диапазона*, нажать *сортировка*.



Расчет необходимых величин

1. В ячейке С1 вводим заголовок *Размах*.
2. В ячейке С2 вводим формулу «=В31-В2». Для этого необходимо в ячейке С2 с клавиатуры ввести знак =, щелкнуть мышкой по ячейке В31, с клавиатуры ввести знак -, щелкнуть мышкой по ячейке В2. После нажатия Enter, в ячейке С2 появится число 8.
3. В ячейке D1 вводим заголовок *число Кл.*
4. В ячейке D2 вводим число классов 4.
5. В ячейку E1 вводим заголовок *Ширина*.
6. В ячейку E2 вводим формулу «=С2/D2». Для этого необходимо в ячейке E2 ввести с клавиатуры знак =, щелкнуть мышкой по ячейке С2, с клавиатуры ввести знак /, щелкнуть мышкой по ячейке D2. После нажатия Enter, в ячейке E2 появится число 2.
7. В ячейку F1 вводим заголовок *Объем*.
8. В ячейку F2 вводим объем выборки 30.

Определение границ интервалов

1. В ячейках E5, E6, E7, E8 вводим заголовки 1кл, 2кл, 3кл, 4кл.
2. В ячейках F4, G4 ввести заголовки нижняя, верхняя.
3. В ячейке F5 ввести минимальное число, т.е. содержимое ячейки В2. Для этого в ячейку F5 ввести формулу «=В2». После нажатия Enter, в ячейке F5 появится число 66.
4. В ячейку G5 вводим формулу «=F5+\$E\$2». Для этого необходимо в ячейке G5 ввести с клавиатуры знак =, щелкнуть мышкой по ячейке G5, с клавиатуры ввести знак +, щелкнуть мышкой по ячейке E2. Для того, чтобы зафиксировать ячейку E2 для процесса копирования, необходимо ее адрес окружить знаками \$. После нажатия Enter, в ячейке G5 появится число 68.
5. В ячейку F6 необходимо перенести данные ячейки G5. Для этого в ячейку F6 необходимо ввести формулу «=G5». После нажатия Enter, в ячейке F6 появится число 68.
6. В ячейки F7, F8 копируем содержимое ячейки F6. Для этого необходимо выделить ячейку F6, подвести указатель мыши к правому нижнему углу ячейки F6, нажать правую кнопку мыши и, не отпуская ее, протянуть до ячейки F8.

7. В ячейки G6-G8 аналогичным способом скопировать ячейку G5. Выполнив эти действия получим:

	E	F	G	H
		нижняя	верхняя	
1кл		66	68	
2кл		68	70	
3кл		70	72	
4кл		72	74	

Расчет логических функций.

Для подсчета частоты попадания в каждый класс необходимо ввести логическую функцию, которая проверяла бы, попадает данное число в данный класс или нет. В качестве такой функции можно использовать логическую функцию ЕСЛИ.

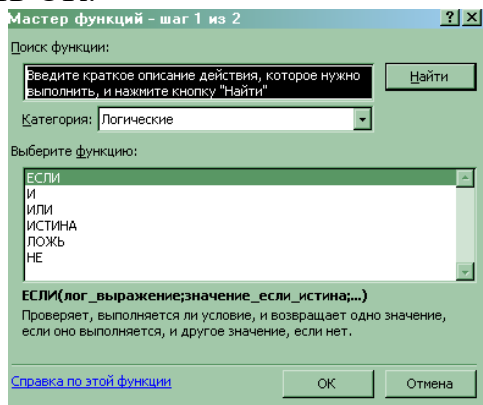
Структура данной функции

ЕСЛИ(лог_выражение; Значение_если_истина;...)

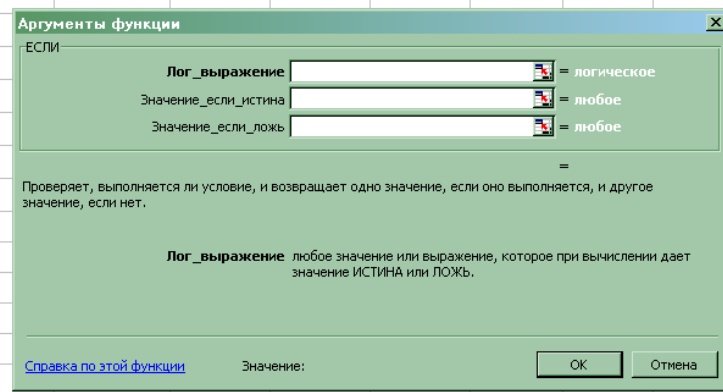
Данная функция проверяет, выполняется ли условие, и возвращает одно значение, если оно выполняется, и другое значение, если нет.

1. В ячейку H1 вводим заголовок 1 класс, I1- 2 класс, J1- 3 класс, K1- 4 класс.

2. В ячейку H2 вводим логическую функцию ЕСЛИ. Для этого в ячейке H2 вводим с клавиатуры знак =, в пункте меню *Вставка* выбрать *Функция*, в категории выбрать *Логические* из предлагаемых функций выбрать ЕСЛИ. нажать ОК.



3. В окне *Аргументы функции* в пункт *Лог_выражение* ввести «B2<\$G\$5», в пункт *Значение_если_истина* 1, в пункт *Значение_если_ложь* 0.



После чего нажать ОК. При этом происходит проверка: если число вариационного ряда меньше, чем последнее число первого класса, то в ячейке Н2 появится число 1, иначе число 0.

4. Скопировать ячейку Н2 в ячейки Н3-Н31. Для этого выделить ячейку Н2, подвести указатель мыши к правому нижнему углу ячейки Н2, протянуть указатель мыши до ячейки Н31.

5. В столбце I рассчитываем частоту попадания во второй класс. Для этого в ячейку I2 вводим формулу: «=ЕСЛИ(Н2=1;0;ЕСЛИ(В2<\$G\$6;1;0))». Первое условие для чисел, которые уже попали в первый класс, выставляет значение 0. Логическая функция ЕСЛИ(В2<\$G\$6;1;0) проверяет условие: если число вариационного ряда меньше, чем последнее число второго класса, то в ячейке I2 появится число 1, иначе число 0.

6. Скопировать ячейку I2 в ячейки I3-I31. Для этого выделить ячейку I2, подвести указатель мыши к правому нижнему углу ячейки I2, протянуть указатель мыши до ячейки I31.

7. В столбце J рассчитываем частоту попадания в третий класс. Для этого в ячейку J2 вводим формулу: «=ЕСЛИ(Н2+I2=1;0;ЕСЛИ(В2<\$G\$7;1;0))». Первое условие для чисел, которые уже попали в первый и второй класс, выставляет значение 0. Логическая функция ЕСЛИ(В2<\$G\$7;1;0) проверяет условие: если число вариационного ряда меньше, чем последнее число третьего класса, то в ячейке J2 появится число 1, иначе число 0.

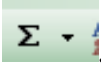
8. Скопировать ячейку J2 в ячейки J3-J31. Для этого выделить ячейку J2, подвести указатель мыши к правому нижнему углу ячейки J2, протянуть указатель мыши до ячейки J31.

9. В столбце K рассчитываем частоту попадания в четвертый класс. Для этого в ячейку K2 вводим формулу: «=ЕСЛИ(Н2+I2+J2=1;0;ЕСЛИ(В2<=\$G\$8;1;0))». Первое условие для чисел, которые уже попали в первый, второй или третий класс, выставляет значение 0. Логическая функция ЕСЛИ(В2<=\$G\$8;1;0) проверяет условие: если число вариационного ряда меньше или равно, чем последнее число четвертого класса, то в ячейке K2 появится число 1, иначе число 0.

10. Скопировать ячейку K2 в ячейки K3-K31. Для этого выделить ячейку K2, подвести указатель мыши к правому нижнему углу ячейки K2, протянуть указатель мыши до ячейки K31.

Подсчет частоты попадания в каждый класс

Для нахождения частоты попадания в каждый класс, необходимо просуммировать значения столбцов H,I,J,K.

1. Выделить ячейки H2-H31.
2. На панели инструментов выбрать кнопку *Автосумма* . После чего в ячейке H32 появится число 4.
3. Выделить ячейки I2-I31.
4. На панели инструментов выбрать кнопку *Автосумма*. После чего в ячейке I32 появится число 5.
5. Аналогичные действия для ячеек J и K приведут к тому, что в ячейках J32 и K32 появятся числа 13 и 8.

Расчет функции плотности вероятности

1. В ячейку A34 вводим заголовок *Частота*.
2. Для удобства расчетов перенесем данные из ячеек H32-K32 в ячейки A35-A38. Для этого в ячейку A35 вводим формулу «=H32», в ячейку A36 – «=I32», в ячейку A37 – «=J32», в ячейку A38 – «=K32».
3. В ячейку B34 вводим заголовок: *Функция*.
4. В ячейке B35 вводим формулу: «=A35/(\$F\$2*\$E\$2)». Для этого необходимо в ячейке B35 с клавиатуры ввести знак =, щелкнуть мышкой по ячейке A35, с клавиатуры ввести знак /, щелкнуть мышкой по ячейке F2, с клавиатуры ввести знак *, щелкнуть мышкой по ячейке E2, с клавиатуры ввести знак), добавить знаки \$. После нажатия Enter, в ячейке B35 появится число 0.666667.
5. Скопировать ячейку B35 в ячейки B36-B38. Для этого необходимо выделить ячейку B35, подвести указатель мыши к правому нижнему углу ячейки B35, протянуть указатель мыши до ячейки B38. Выделяем ячейки B35-B38. С помощью пункта меню *Формат/ячейки* выбрать числовой формат *Числовой*, число десятичных знаков 2. При этом появится таблица

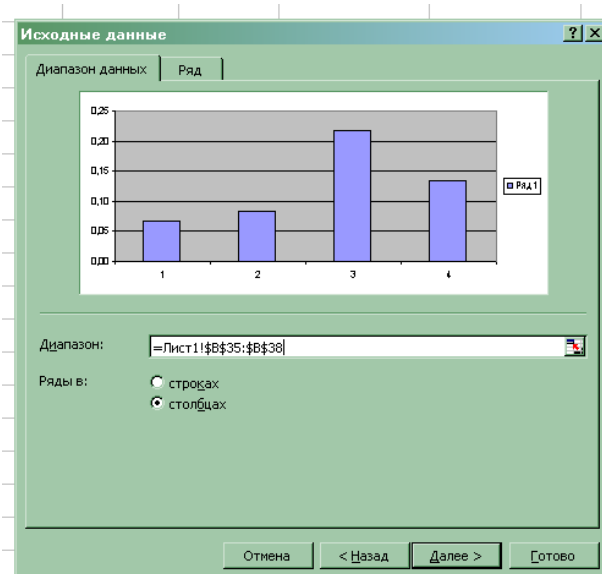
33			
34	частота	функция	
35	4	0,07	
36	5	0,08	
37	13	0,22	
38	8	0,13	
39			

Построение гистограммы

1. Для более наглядного построения гистограммы, перенесем границы интервалов в такой форме:

39			
40			
41	частота	функция	границы
42	4	0,07	66
43	5	0,08	70
44	13	0,22	72
45	8	0,13	74
46			
47			

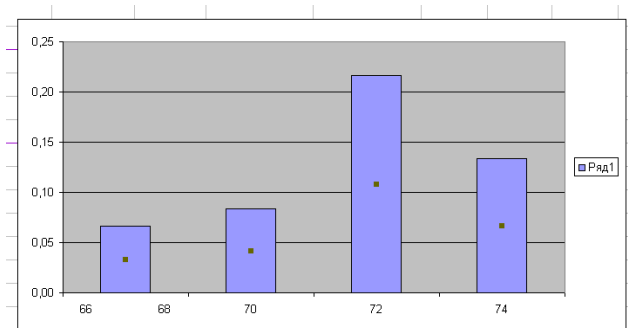
1. На панели инструментов выбрать кнопку *Мастер диаграмм*.
В появившемся окне выбрать тип *Гистограмма*, затем нажать *Далее*.
2. Появится окно *Мастер диаграмм*.



В поле *Диапазон* удалить появившуюся запись, щелкнуть по ячейке B42, протянуть мышку до ячейки B45.

3. Зайти во вкладку *Ряд*. Щелкнуть мышкой в поле *Подписи оси X*, затем щелкнуть мышкой по ячейке C42, протянуть указатель мыши до ячейки C45. Затем нажать *Далее* и *Готово*.

4. Щелкнуть мышкой по одному из прямоугольников гистограммы так, чтобы в центре каждого из прямоугольников появилась точка. Щелкнуть правой кнопкой мыши и в появившемся окне выбрать пункт *Формат рядов данных*.



5. В появившемся окне выбрать вкладку *Параметры*. Затем уменьшить ширину зазора до 0, нажать *Ок*. При этом получим гистограмму