Кафедра информатики и компьютерных технологий

ВВЕДЕНИЕ В ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

АНАЛИЗ ДАННЫХ. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Методические указания для выполнения курсовой работы для студентов бакалавриата направления подготовки 050306 «Экология и природопользование»

САНКТ-ПЕТЕРБУРГ
2023

УДК 519.86:622.3.012 (075.83)

АНАЛИЗ ДАННЫХ. Описательная статистика: Методические указания для выполнения курсовой работы / Сост.: *В.В. Беляев, Т.Р. Косовцева.* СПб, 2023. 32 с.

Методические указания содержат теоретические сведения по выполнению курсовой работы по дисциплине «Введение в информационные технологии». Приведены примеры выполнения заданий по некоторым разделам математической статистики: построение вариационного и интервального рядов, вычисление выборочных характеристик (описательных статистик), получение статистических оценок и определение их свойств.

Все расчеты выполнены с использованием электронных таблиц MS Excel, в том числе с применением надстройки «Пакет анализа».

Методические указания предназначены для студентов бакалавриата направления подготовки 05.03.06 «Экология и природопользование» дневной формы обучения.

Табл.5. Рис.21. Библиогр.: 3 назв.

Научный редактор: *доц. Журов Г.Н.*

ВВЕДЕНИЕ

Целью курсовой работы по дисциплине «Введение в информационные технологии» является углубление знаний и закрепление навыков полученных студентами при изучении дисциплины «Введение в информационные технологии» на I курсе.

В задачах экологии и природопользования часто имеем совокупность наблюдений, на основе которых нужно сделать какие-либо выводы. Часто подобных наблюдений много, и возникает задача их компактного описания с использованием различных параметров. Эти параметры могут быть оценены с помощью методов описательной статистики.

Описательная одномерная статистика обеспечивает простой путь для организации и систематизации выборочных данных. Статистическое описание выборочных данных должно предварять любой статистический анализ.

Основная идея *выборочного метода* – получение достоверной информации о генеральной совокупности по заданной выборочной совокупности.

Курсовая работа предполагает выполнение каждым студентом обработку экспериментальных данных методом описательной статистики, используя навыки работы в табличном процессоре MS Excel и математическом пакете MathCad.

Методические указания содержат всю необходимую информацию для выполнения курсовой работы.

- сведения об основных этапах работы, начиная от формализации задач и кончая защитой отчета о выполненной работе;
- постановку предлагаемых для решения задач и разработку математической модели;
- указания по вводу расчетных формул и по способу формализации данных с соответствующими примерами;
- варианты задач;
- рекомендательный библиографический список.

1 ПОРЯДОК ВЫПОЛНЕНИЯ КУРСОВОЙ РАБОТЫ

1). Согласование темы с руководителем работы.
2). Изучение литературы по теме.

3). Повторение, углубленное изучение разделов учебников, относящихся к выбранной теме;
4). Ознакомление с литературой, рекомендуемой настоящими методическими указаниями, конспектирование и цитирование необходимых для решения поставленной задачи теоретических положений;
5). Самостоятельный подбор дополнительной литературы;
6). Подбор справочного материала.
7). Составление текста заданий в соответствии с номером варианта.
8). Формализация исходной информации, выбор методов решения.
9). Выполнение расчетов в табличном процессоре MS Excel.
10). Выполнение расчетов в системе MathCad.
11). Написание пояснительной записки (отчета о работе)..
12). Сдача работы на проверку ее руководителю, доработка текста, графики.
13). Защита курсовой работы.

Выдача заданий по курсовой работе производится не позднее, чем через две недели после начала занятий. Во время выдачи заданий объявляются сроки выполнения студентом отдельных этапов, назначается дата сдачи отчета на проверку и дата защиты работы (календарный план).

При выставлении оценки по курсовой работе учитываются правильность решения задачи, качество отчета, оригинальность и творческий подход к выбору методов решения, а также своевременность выполнения всей работы и отдельных ее этапов.

Студент обязан не менее одного раза в месяц информировать руководителя курсовой работы о выполненных этапах.

2. ТРЕБОВАНИЯ К ОТЧЕТУ ПО РАБОТЕ

Отчет по курсовой работе (пояснительная записка) должен содержать следующие разделы:

- титульный лист;
- задание по курсовой работе;
- аннотация;

- оглавление;
- введение;
- теоретические сведения;
- постановка задачи;
- исходные данные;
- подробное описание решения задачи при использовании табличного процессора MS Excel;
- выполнение расчетов в системе MathCad;
- результаты расчетов в виде графиков и таблиц;
- анализ решения задачи, выводы;
- библиографический список;
- приложения.

На титульном листе указывается официальное название института, вид работы, наименование кафедры и название дисциплины, тема курсовой работы, фамилия и инициалы студента, шифр группы, дата оформления отчета, должность, фамилия и инициалы руководителя работы, место для выставления оценки.

В аннотации приводятся краткие сведения о содержании работы (на русском и иностранном языках), количество страниц рисунков, таблиц в отчете, и количество наименований в библиографическом списке.

Введение должно содержать информацию о наиболее часто используемых программных средствах при решении задач из предметной области (математические и статистические пакеты прикладных программ).

Теоретические сведения должны содержать информацию, необходимую для решения задачи в общем виде. При указании формул следует разъяснить смысл всех величин, входящих в них.

Текст каждой задачи составляется студентом с учетом постановки задачи и конкретных данных, соответствующих номеру студента в списке группы.

Решение задачи с помощью табличного процессора MS Excel должно демонстрировать этапы расчета с необходимыми для их понимания комментариями. В отчете приводятся фрагменты рабочих

листов в режиме отображения данных и в режиме отображения формул.

Результаты вычислений в MS Excel следует использовать в качестве контроля для проверки правильности решения в математическим пакете MathCad (SMath Studio). Расчеты в системе MathCad выполняются с использованием встроенных функций. Все расчеты в документе MathCad (SMath Studio) необходимо прокомментировать.

Формализация задачи предполагает, что должны быть рассмотрены вопросы: в какой форме представить исходные данные для их ввода в компьютер, какие формулы и в какой последовательности следует применить для получения промежуточных и окончательных результатов, какова точность вычисления всех параметров и правила их округления. Конкретные рекомендации для каждой задачи имеются в настоящих разделах методических указаний.

3 ИЗУЧЕНИЕ БАЗОВЫХ ПОНЯТИЙ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

3.1 БАЗОВЫЕ ПОНЯТИЯ

При исследовании реальных экономических процессов приходится обрабатывать большие объёмы статистических данных, которые по своей сути являются случайными величинами (СВ). На практике количество реализаций СВ ограничено, что не позволяет применять напрямую теоретические методы анализа. Поэтому при обработке данных в первую очередь используют методы и модели математической статистики, позволяющие получить необходимые знания об исследуемом объекте.

3.1.1 ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА

Статистической совокупностью называется множество предметов или явлений, объединённых в нечто целое и однородное по некоторым определенным признакам. Отдельные элементы, входящие в совокупность, называются *членами статистической совокупности*, а общее число членов совокупности – её *объёмом*.

Изменение признака при переходе от одного члена совокупности к другому называют его *вариацией*, а значение

признака у отдельного члена статистической совокупности – его *вариантой*.

Выборочной совокупностью (или выборкой) называется совокупность случайно отобранных однородных элементов. *Генеральной совокупностью* называется совокупность всех однородных элементов, из которых произведена выборка.

Выборочная и генеральная совокупности, как правило, различаются объемами. Выборка называется *репрезентативной*, если она достаточно хорошо представляет исследуемый признак генеральной совокупности. Для обеспечения репрезентативности выборки применяют следующие способы отбора: *простой отбор* (последовательно отбираются последовательно случайно попавшиеся объекты), *типический отбор* (объекты отбираются пропорционально представительству различных типов объектов в генеральной совокупности), *случайный отбор*, например, с помощью таблицы случайных чисел, и т.д.

Одной из основных задач статистического анализа является получение по полученной выборке достоверных сведений об интересующих исследователя свойствах и параметрах генеральной совокупности.

Основным типом значений переменных в математической статистике являются количественные переменные.

3.1.2 ВЫЧИСЛЕНИЕ ВЫБОРОЧНЫХ ХАРАКТЕРИСТИК

Значения количественных переменных являются числовыми, могут быть упорядочены и для них имеют смысл различные вычисления (например, вычисление среднего значения). На обработку количественных переменных ориентировано подавляющее большинство статистических методов.

Первый раздел математической статистики – *описательная статистика* – предназначен для представления исследуемых данных в удобном виде и для получения информации о них в терминах математической статистики и теории вероятностей. Для этого используются описательные или дескриптивные характеристики:

- минимум,
- максимум,
- размах,

- среднее,
- дисперсия,
- стандартное отклонение,
- медиана,
- квартили,
- мода.

При анализе конкретного показателя *X* все элементы выборки $x_1, x_2, \ldots, x_n$ объемом *n* обычно упорядочивают по неубыванию:

$$x_1 \le x_2 \le \text{K} \le x_n .$$

Выборка, упорядоченная по возрастанию наблюдаемых значений, называется *вариационным рядом.*

Разность между максимальным и минимальным значениями ряда *X* называется *размахом* выборки.

Если значение x_i встречается в выборке n_i раз, то число n_i называется *частотой (частостью)* значения x_i, а величина

$$m_i = \frac{n_i}{n}$$

- *относительной частотой* значения x_i.

Генеральная совокупность может быть характеризована множеством **параметров**. К их числу можно отнести среднее значение, дисперсию и ряд других. Определение основных параметров приведено ниже.

Пусть объем генеральной совокупности равен *N*, как правило, *n*<<*N*.

Тогда величина $\bar{x}_\Gamma = \frac{1}{N}\sum_{i=1}^{N} x_i$ является *генеральной средней*.

Генеральной дисперсией является величина

$$D_\Gamma = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}_\Gamma\right)^2 .$$

Генеральным средним квадратическим отклонением является величина $\sigma_\Gamma = \sqrt{D_\Gamma}$.

Как правило, все элементы генеральной совокупности не известны, поэтому точные значения **параметров** найти невозможно,

чаще всего приходится работать с выборками из генеральной совокупности. Тогда соответствующие **оценки параметров** можно найти по следующим формулам.

- *выборочное среднее:*

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (3.1)$$

- выборочная дисперсия:

$$D_в = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_в\right)^2 \qquad (3.2)$$

$$D_{испр ав} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}_в\right)^2 \qquad (3.3)$$

Второй вариант формулы целесообразно использовать при $n<30$.

- выборочное среднее квадратическое отклонение:

$$\sigma_в = \sqrt{D_в} \qquad (3.4)$$

Выборочное среднее является **о ц е н к о й** генерального среднего, выборочная дисперсия и дисперсия исправленная являются оценками генеральной дисперсии.

Другие **оценки параметров** приведены ниже, служат для оценки соответствующих **параметров** генеральной совокупности.

Мода MoX – это наиболее часто встречающееся значение признака в данном ряду распределения. Для дискретных вариационных рядов мода определяется как значение признака с наибольшей частотой. В случае непрерывной вариации мода может быть определена как значение признака, которому отвечает наибольшая плотность распределения частости.

Медианой $M_e X$ называется значение признака, относительно которого статистическая совокупность делится на две равные по объему части. При этом в одной из них содержатся члены, у которых значения признака не больше, а в другой – члены со значениями

признака не меньше, чем $M_e X$. Другими словами, *медианой* называется число, разделяющее выборку пополам: 50% элементов меньше медианы, а 50% -больше медианы.

Медиана рассчитывается по-разному в дискретных и интервальных рядах.

Если в дискретном ряду распределения нечетное число уровней, то медианой будет серединное значение упорядоченного ряда признака, т.е. это элемент с номером $\frac{n+1}{2}$ вариационного ряда.

Если ряд распределения дискретный и состоит из четного числа членов, то медиана определяется как средняя величина из двух серединных значений вариационного ряда.

Квартили – это показатели, которые чаще всего используются для оценки распределения данных при описании свойств больших числовых выборок. В то время, как медиана разделяет упорядоченный массив пополам, квартили разбивают упорядоченный массив данных на четыре части.

Первый квартиль Q_1 – это число, разделяющее выборку на две части: 25% элементов меньше, а 75% - больше первого квартиля (3.5).

$$Q_1 = \frac{n+1}{4} \tag{3.5}$$

Третий квартиль Q_3 – это число, разделяющее выборку на две части: 75% элементов меньше, а 25% - больше третьего квартиля(3.6).

$$Q_3 = \frac{3(n+1)}{4} \tag{3.6}$$

Для вычисления квартилей применяются следующие правила.

1. Если индекс квартиля задается целым числом, значением квартиля считается элемент выборки с указанным индексом.
2. Если индекс квартиля задается величиной, представляющей собой среднее значение, вычисляемое по двум целым числам, квартиль равен среднему арифметическому, вычисленному по элементам, индексы которых равны эти двум числам.

3. Если индекс квартиля задается числом, которое не является целым и не кратно ½, он просто округляется до ближайшего целого. Квартилем является элемент выборки с указанным индексом.

Асимметрия – это свойство распределения выборки, которое характеризует несимметричность распределения СВ. На практике симметричные распределения встречаются редко, и чтобы выявить и оценить степень асимметрии, вводят следующую меру:

$$A_s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3}. \qquad (3.7)$$

Пределы значений A_s от $-\infty$ до $+\infty$

При $A_s = 0$ распределение симметрично: $MoX = \bar{x}$.

При положительной асимметрии $MoX < \bar{x}$;

При отрицательной асимметрии $MoX > \bar{x}$.

Эксцесс – это мера крутости кривой распределения. *Эксцесс* вычисляется по формуле:

$$E_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sigma^4} - 3. \qquad (3.8)$$

Значения E_k лежат в открытом интервале $[-3,+\infty[$. Если $E_k > 0$, то кривая распределения имеет более острую вершину, чем нормальное распределение с параметрами $m = \bar{x}_в$ и $\sigma = \sqrt{D_в}$, и распределение называется островершинным.

Если $E_k < 0$, то кривая распределения имеет более плоскую вершину, чем нормальное, и распределение называется плосковершинным.

Для нормального распределения $A_s = 0$, $E_k = 0$.

3.1.3. ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Эмпирической функцией распределения называется следующая функция:

$$F(x)=\begin{cases}0, & при\ x \le x_1\\ \dfrac{i}{n}, & при\ x_i < x \le x_{i+1}\\ 1, & при\ x > x_n\end{cases}. \qquad (3.9)$$

Эта формула справедлива, когда все x_k различны. Знание функции распределения позволяет вычислить различные характеристики исследуемого признака.

3.1.4. ИНТЕРВАЛЬНЫЙ ВАРИАЦИОННЫЙ РЯД

Чтобы получить первое впечатление о распределении генеральной совокупности, необходимо провести некоторую обработку выборочных данных. Простейшей операцией является **построение интервального ряда**.

Если произвести группировку вариант по интервалам изменения признака (*интервальная группировка*) и результат представить рядом интервалов вариант, расположенных в порядке их возрастания, и рядом соответствующих частот, то получим *интервальный вариационный ряд*.

Под *частотой значения признака* или *интервала* понимают число членов совокупности, варианты которых лежат в данном интервале. Отношение частоты к объему совокупности называется относительной *частотой* или *частостью*.

Число равных интервалов *k*, на которые следует разбить весь диапазон значений признака $X\left\lfloor x_{min,}\ x_{max}\right\rfloor$, может быть найдено по формуле:

$$k = log_2 n + 1, \qquad (3.10)$$

где *n* – объем статистической совокупности.

Число интервалов должно быть не меньше 8-10 и не больше 20-25.

Размах выборки определяется по формуле:

$$\Delta = x_{max} - x_{min}, \qquad (3.11)$$

а длина интервала - по формуле:

$$h = \frac{\Delta}{k}. \qquad (3.12)$$

Формулы 3.10 и 3.12 дают **оценочное значение** количества интервалов и их размеров. При практическом построении рекомендуется брать значения k и h, которые соответствуют здравому смыслу.

В дальнейшем будем использовать следующие обозначения:
a_i, b_i - левая и правая границы i-го интервала соответственно;
x_i – середина этого интервала;
m_i – частота интервала.

3.1.5 ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ИНТЕРВАЛЬНЫХ ВАРИАЦИОННЫХ РЯДОВ

Для наглядного представления статистического распределения пользуются графическим изображением интервальных вариационных рядов. К числу таких графических изображений относятся гистограмма, полигон, кумулята.

1). Построение гистограммы.

Для построения гистограммы нужно составить таблицу, в которой необходимо указать границы интервалов, найти их середины и частоту значений признака для каждого интервала.

Пример

Пусть объем выборки равен n=60; x_1=-2,18; x_{60}=12,04; h=2; k=9. В табл.3.1 представлен соответствующий интервальный вариационный ряд.

Таблица 3.1

№ (разряд)	Границы интервала		Середина интервала *x[i]*	Частота *mi*
	левая *a[i]*	правая *b[i]*		
1	-4	-2	-3,00	1
2	-2	0	-1,00	3
3	0	2	1,00	6
4	2	4	3,00	12
5	4	6	5,00	20
6	6	8	7,00	8
7	8	10	9,00	8
8	10	12	11,00	1
9	12	14	13,00	1
sum				60

По оси абсцисс откладывают интервалы значений признака, и на каждом из них, как на основании, строят прямоугольник с высотой, пропорциональной частоте интервала.

Гистограмма, построенная с помощью *Мастера диаграмм* программы MS Excel, приводится на рис. 3.1.

2). Построение полигона

Для построения *полигона* на оси абсцисс откладывают интервалы значений признака, в серединах интервалов восстанавливают перпендикуляры, длины которых пропорциональны соответствующим частотам, затем концы соседних перпендикуляров соединяют отрезками прямых, а концы крайних перпендикуляров соединяют с серединами соседних интервалов, частоты которых равны нулю. В результате получим замкнутую фигуру в виде многоугольника.

Полигон для интервального ряда приведен на рис.3.2.

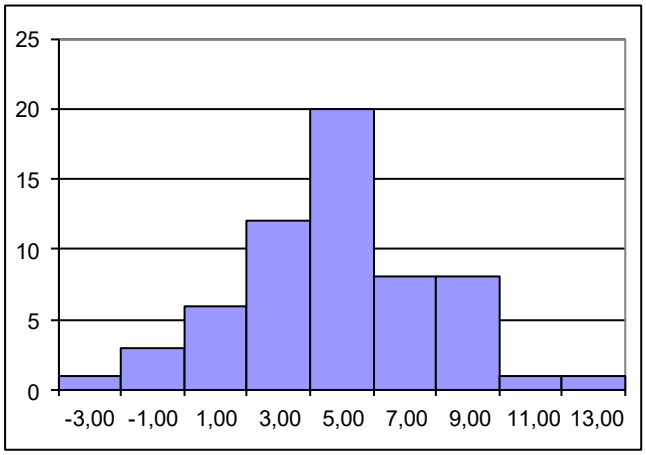

Рис.3.1 Гистограмма, построенная с помощью Мастера диаграмм MS Excel

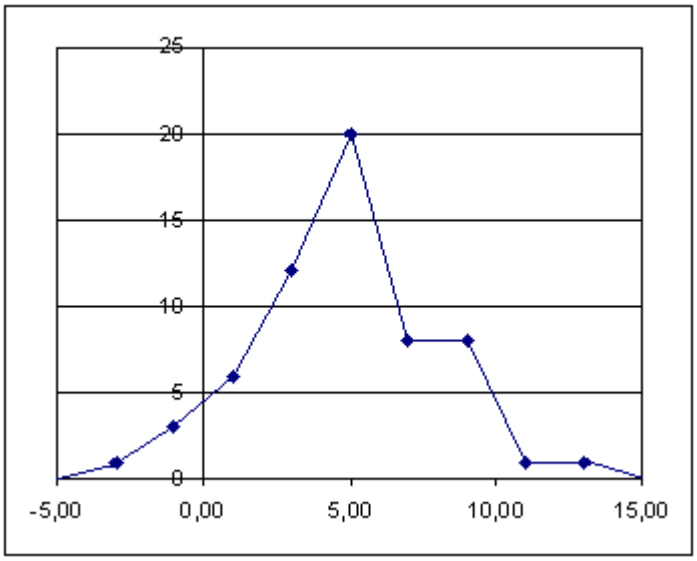

Рис.3.2. Полигон интервального ряда

3). Построение кумуляты

Накопленной частостью (частотой) в точке x называют суммарную частость (частоту) членов статистической совокупности со значениями признака меньшими, чем x.

Если в вариационном ряду вместо частот или частостей записать соответственно накопленные частоты или частости, то получится *кумулятивный ряд*. Для графического построения кумулятивных рядов пользуются кумулятами.

Кумулята строится следующим образом: на оси абсцисс отмечают точки, соответствующие границам интервалов или значениям признака. В каждой такой точке восстанавливают перпендикуляр, длина которого пропорциональна накопленной частоте. Концы соседних перпендикуляров соединяют отрезками. Полученная ломаная линия называется *кумулятой*.

Эмпирическая функция распределения отличается от кумуляты только масштабом.

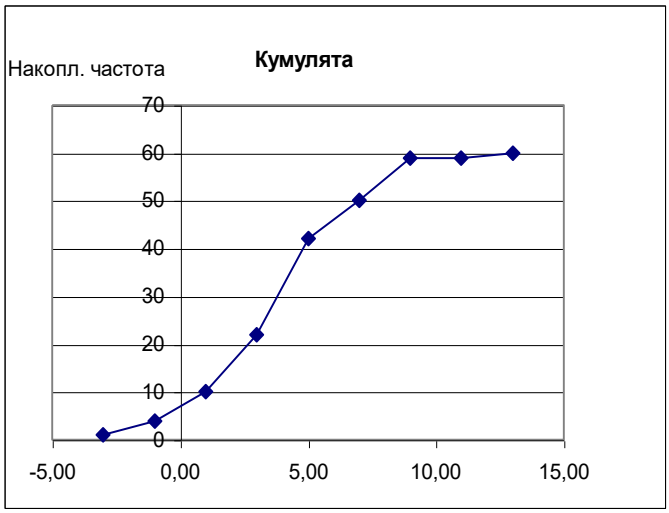

Рис.3.3. Кумулята интервального ряда

Используя полученный интервальный ряд, можно вычислить **все оценки парметров**, полагая, что все варианты выборки, лежащие внутри *i*-го интервала, принимают значения равные x_i с частотой m_i **.** Тогда выборочное среднее $\bar{x}_в$ и выборочная дисперсия $D_в$ вычисляются по формулам:

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{k} m_i x_i \,, \qquad (3.13)$$

$$D_в = \frac{1}{n}\sum_{i=1}^{k} m_i \left(x_i - \bar{x}_в\right)^2 . \qquad (3.14)$$

Если $a_j - b_j$ - *модальный интервал*, т.е. интервал, которому соответствует наибольшая частота m_j , а интервалы вариационного ряда имеют постоянную ширину *h*, то мода признака вычисляется по формуле

$$MoX = a_j + h \cdot \frac{m_j - m_{j-1}}{(m_j - m_{j-1}) + (m_j - m_{j+1})}, \qquad (3.15)$$

где m_{j-1}, m_{j+1} – частоты интервалов, предшествующих модальному и следующего за модальным, соответственно.

Для интервального распределения сначала находят так называемый *медианный интервал* $a_s - b_s$, номер которого вычисляют из неравенств

$$\gamma(a_s) \le 0{,}5; \;\; \gamma(b_s) > 0{,}5; \qquad (3.16)$$

где $\gamma(x)$ - накопленная частота в точке *x*. В предположении, что в медианном интервале признак распределен равномерно, медиана признака *X* определяется по формуле:

$$MeX = a_s + h \cdot \frac{\frac{n}{2} - \gamma(a_s)}{m_s}, \qquad (3.17)$$

где h – ширина интервала *c* номером *s*; m_s – частота этого интервала.

Необходимо отметить, что оценки, полученные с помощью интервального ряда являются менее точными по сравнению с оценками полученными по исходной выборке (вариационному ряду). Применение этих формул целесообразно, когда исходные данные не доступны.

3.2. ДИАГРАММА ТИПА “ЯЩИК С УСАМИ”

3.2.1. ОБЩИЕ СВЕДЕНИЯ

Диаграмма типа “ящик с усами” изображает важные характеристики описательной статистики на одном компактном рисунке. Она предложена Джоном Тьюки (John Tukey) в 1977 г. в основополагающей книге *Exploratory Data Analysis*. Диаграмма типа “ящик с усами” отображает следующие характеристики СВ:

1. Первый квартиль, медиана, третий квартиль и интерквантильный диапазон;
2. Минимальное и максимальное значения;
3. Умеренные и экстремальные выбросы.

Диаграмма типа “ящик с усами” дает хорошее визуальное представление изменчивости данных, а также асимметрии распределения. Типичный вид диаграммы типа “ящик с усами” приведен на рис.3.4.

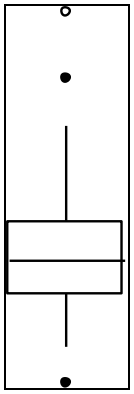

Рис. 3.4. Диаграмма типа “ящик с усами”

3.2.2. ИНТЕРКВАРТИЛЬ

Первый компонент диаграммы типа “ящик с усами” называется *интерквартиль* или *интерквартильный диапазон* (*interquartile range — IQR*), который простирается от первого до третьего квартиля.

Интерквартиль (*IQR)* - одна из мер разброса или рассеяния данных. Он равен разности между верхним и нижним (первым и третьим) квартилями. Другими словами *IQR* - это ширина интервала, содержащего средние 50% выборки. Таким образом, чем меньше *IQR,* тем меньше рассеяние. Положительной чертой этого показателя является его устойчивость (робастность), т.е. на него слабо влияют выбросы.

Пример.

Пусть дана выборка (уже в виде вариационного ряда):

2 3 4 5 6 6 6 7 7 8 9.

Ее верхний квартиль равен 7, ее нижний квартиль равен 4, следовательно, IQR равняется 7 - 4 = 3.

Для создания интерквартиля строят прямоугольник (“ящик”) от первого до третьего квартиля. Внутри ящика проводят горизонтальную линию на уровне медианы (второго квартиля) (рис. 3.5).

3.2.3. ОГРАЖДЕНИЯ

После построения интерквартильного диапазона можно приступать к вычислению внутреннего и внешнего ограждений. *Внутренние ограждения* (*inner fences*) располагаются в области, большей третьего квартиля + $1{,}5\times IQR$ или меньшей первого квартиля $-1{,}5\times IQR$. *Внешние ограждения* (*outer fences*) располагаются в области большей третьей квартили $+3\times IQR$ или меньшей первой квартили $-3\times IQR$ (рис. 3.5).

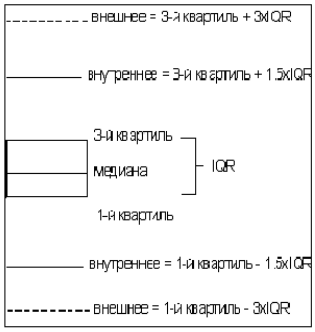

Рис.3.5. Расположение ограждений при построении диаграммы «ящик с усами»

Замечание. *Диаграммы на рис.3.5-3.8 нарисованы без* точного соответствия масштабу.

3.2.3. ВЫБРОСЫ

Все значения выборки, которые лежат в промежутке между внутренним и внешним ограждениями, называются *умеренными выбросами* (*moderate outlier*) и обозначаются символами ●. Все значения, которые лежат за пределами внешних ограждений, называются *экстремальными выбросами* (*extreme outlier*) и обозначаются символами ○ (рис. 3.6).

Рис.3.6. Выбросы при построении диаграммы «ящик с усами»

3.2.4. УСЫ

Усы - вертикальные линии, проведенные от «ящика» до максимального и минимального значения СВ внутри внутреннего ограждения (рис. 3.7), такие значения *не* считаются выбросами.

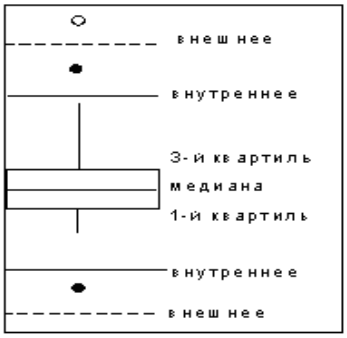

Рис.3.7. Расположение усов при построении диаграммы «ящик с усами»

3.2.5. ОКОНЧАТЕЛЬНЫЙ ВИД ДИАГРАММЫ

Обычно в окончательном виде статистической диаграммы типа “ящик с усами” внутреннее и внешнее ограждения не отображаются. Обычно эта диаграмма выглядит так, как показано на рис.3.8. Как видите, в этих данных имеются три выброса, причем один из них является экстремальным, а распределение в целом асимметрично.

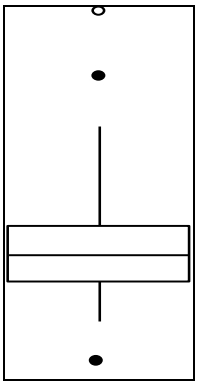

Рис.3.8. Окончательный вид статистической диаграммы типа “ящик с усами”

3.3. НАДСТРОЙКА «ПАКЕТ АНАЛИЗА» MS EXCEL

Надстройка (модуль) «Пакет анализа» MS Excel предназначен для выполнения базовых операций статистического анализа. Полученные с его помощью результаты не обновляются при изменении исходных данных, поэтому после их изменения для обновления результатов требуется снова выполнить соответствующую команду.

Для активизации надстройки **Пакет анализа** выберите пункт меню **Данные →Анализ данных.** Если этот пункт меню недоступен, загрузите **Пакет анализа** (кнопка Office→ **Настройка** MS Excel → **Надстройки)**.

3.3.1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Это средство анализа (рис.3.10) служит для создания таблицы с точечными оценками одномерной выборки

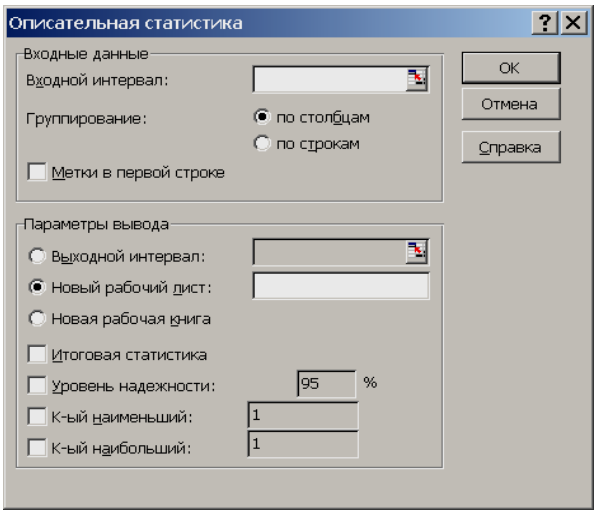

Рис.3.10 Диалоговое окно «Описательная статистика»

Раздел *Входные данные*

Поле **Входной интервал** используется для ввода диапазона смежных ячеек с анализируемыми данными.

Группа переключателей **Группирование** используется для указания способа расположения анализируемых данных по столбцам или по строкам.

Флажок **Метки в первой строке** устанавливают для обозначения того, что первая строка анализируемых данных содержит заголовки столбцов.

Раздел *Параметры вывода*

Выходной интервал - переключатель, используемый для указания начальной ячейки в верхнем левом углу диапазона ячеек, в которых будут располагаться полученные результаты.

Переключатель **Новый рабочий лист** используется для указания того, что результаты будут располагаться на новом рабочем листе с указанным именем.

Флажок **Итоговая статистика** используется для вывода статистических параметров.

3.3.2. РАНГ И ПЕРСЕНТИЛЬ

Инструмент «Ранг и персентиль» - средство анализа, которое используется для вывода таблицы, содержащей порядковый и процентный ранги для каждого значения в наборе данных (рис.3.11).

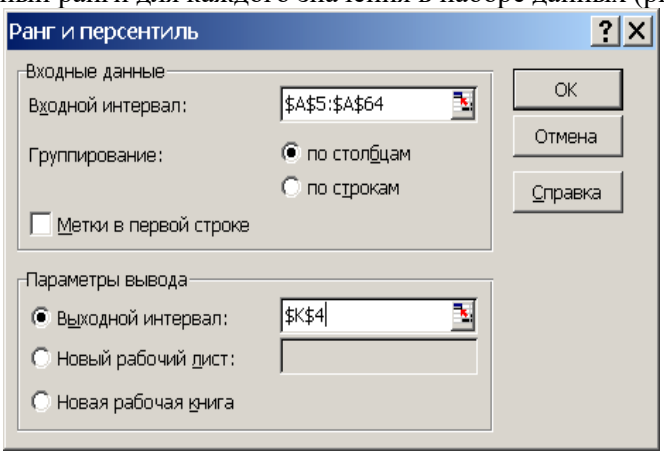

Рис.3.11. Диалоговое окно «Ранг и персентиль»

Данная процедура может быть применена для анализа относительного взаиморасположения данных в наборе и для приближенного построения функции распределения.

3.3.3. ГИСТОГРАММА

«Гистограмма» - один из инструментов *Пакета анализа*. Используется для вычисления выборочных и интегральных частот попадания данных в указанные интервалы значений (рис.3.12).

Параметры диалогового окна "Гистограмма":

Входной интервал. Используют для ввода диапазона смежных ячеек с исследуемыми данными.

Интервал карманов (необязательный). Используют для ввода диапазона ячеек и необязательного набора граничных значений, определяющих отрезки (карманы). Эти значения должны быть введены в возрастающем порядке.

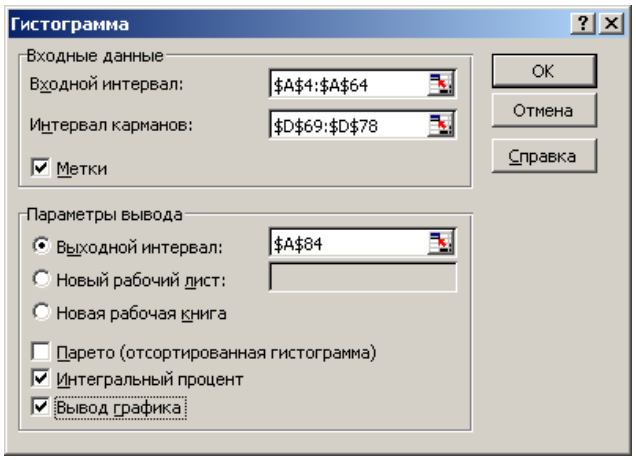

Рис.3.12 Диалоговое окно «Гистограмма»

В Microsoft Excel вычисляется число попаданий данных в интервал, ограниченный началом отрезка и соседним большим по порядку, если такой есть. При этом в интервал включаются значения, принадлежащие нижней границе отрезка и не включаются значения, соответствующие верхней границе.

Если диапазон карманов не был введен, то набор отрезков, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.

Метки. Флажок устанавливают, когда первая строка анализируемых данных содержит заголовки столбцов. Если заголовки отсутствуют; в этом случае подходящие названия для данных выходного диапазона будут созданы автоматически.

Выходной интервал. Используют для указания ссылки на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные.

ЗАДАНИЕ

Из генеральной совокупности извлечена выборка объема n. Изучить распределение непрерывного признака X некоторой генеральной совокупности.

Требуется:

1. Построить вариационный ряд.
2. Найти точечные оценки математического ожидания, дисперсии, среднего квадратического отклонения, моды, медианы, размаха, асимметрии и эксцесса. Вычисления провести в MS Excel следующими способами:
   1) по формулам ;
   2) с помощью стандартных функций Мастера функций;
   3) с помощью надстройки *Пакет анализа*.
3. Вычислить первый, второй и третий квартили, используя встроенную функцию КВАРТИЛЬ();
4. Вычислить 5-ый, 50-ый и 95-ый персентили, используя встроенную функцию ПЕРСЕНТИЛЬ().
5. Построить (приближенно) эмпирическую функцию распределения данного вариационного ряда. С помощью графика этой функции проиллюстрировать результаты, полученные в п.3 и 4.
6. Построить диаграмму типа «Ящик с усами».
7. Построить интервальный вариационный ряд.
8. Построить полигон, гистограмму, кумуляту и эмпирическую функцию распределения для полученного интервального вариационного ряда.
9. Найти точечные оценки числовых характеристик m_X, D_X, σ_X используя интервальный ряд. Сравнить результаты с п.2 (объяснить различия).
10. Построить интервал, содержащий ***K***-% центрального признака по функции распределения, используя встроенную функцию Персентиль (по вариантам).
11. Расчеты выполнить в математическом пакете MathCad (SMath Studio).

ПРИМЕР РЕШЕНИЯ ЗАДАЧИ В MS EXCEL

Исходные данные приведены в виде таблиц Excel в интервале ячеек A5:A64 (рис.3.13).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
3	Значения	признака							"Пакет Анализа –		Ранг и перцентиль"			
4	X										*Точка*	*Столбец1*	*Ранг*	*Процент*
5	-2,18										60	12,04	1	100,00%
6	-0,57							2-й способ			59	9,9	2	98,30%
7	-0,36			1-й способ -				«Пакет Анализа –			58	9,55	3	96,60%
8	-0,3			встроенные ф. Excel				Описательные статистики»			57	9,28	4	94,90%
9	0,43								*X*		56	9,09	5	93,20%
10	0,85										55	9,06	6	91,50%
11	1,33			Среднее		4,64583		Среднее	4,645833333		54	8,44	7	89,80%
12	1,42							Стандартна	0,375927801		53	8,23	8	88,10%
13	1,52			Медиана		4,28		Медиана	4,28		52	8,17	9	86,40%
14	1,58			Мода		4,05		Мода	4,05		51	8,04	10	84,70%
15	2,06			Станд.откл.		2,91192		Стандартно	2,911924224		50	7,48	11	83,00%
16	2,09			Дисперсия		8,4793		Дисперсия	8,479302684		49	7,38	12	81,30%
17	2,27			Эксцесс		-0,0963		Эксцесс	-0,096252528		48	7,21	13	79,60%
18	2,36			Асимметричность		0,11972		Асимметри	0,119723303		47	7,09	14	77,90%
19	2,38			Интервал				Интервал	14,22		46	6,75	15	76,20%
20	2,66			Минимум	-2,18	-2,18		Минимум	-2,18		45	6,23	16	74,50%
21	2,98			Максимум	12,04	12,04		Максимум	12,04		44	6,22	17	72,80%
22	3,54			Сумма				Сумма	278,75		43	6,2	18	71,10%
23	3,65			Объем выборки		60		Счет	60		42	5,92	19	69,40%
49	6,23				Квартили			Персентили			16	2,66	45	25,40%
50	6,75				0	-2,18	мин	0%	-2,18	мин	15	2,38	46	23,70%
51	7,09				1	2,59		5%	-0,303		14	2,36	47	22,00%
52	7,21				2	4,28		50%	4,28		13	2,27	48	20,30%
53	7,38				3	6,36		95%	9,2935		12	2,09	49	18,60%
54	7,48				4	12,04	макс	100%	12,04	макс	11	2,06	50	16,90%
55	8,04										10	1,58	51	15,20%
56	8,17										9	1,52	52	13,50%
57	8,23										8	1,42	53	11,80%
58	8,44										7	1,33	54	10,10%
59	9,06										6	0,85	55	8,40%
60	9,09										5	0,43	56	6,70%
61	9,28										4	-0,3	57	5,00%
62	9,55										3	-0,36	58	3,30%
63	9,9										2	-0,57	59	1,60%
64	12,04										1	-2,18	60	,00%

Рис.3.13. Рабочий лист MS Excel в режиме отображения данных.

Для упорядочивания признака *X* по возрастанию следует воспользоваться командой **Данные→Сортировка**, результат сортировки расположить в этом же интервале.

Для выполнения *пункта 2* задания точечные оценки можно найти двумя способами:

1) с помощью встроенных функций: СРЗНАЧ(), ДИСП(), МЕДИАНА() и т.д. (рис.3.13 и рис.3.14);

2) с помощью надстройки Excel *«Пакет Анализа – Описательные статистики».* Для этого следует заполнить диалоговое окно, приведенное на рис.3.11 и получить результат в интервале ячеек H8:I20. (рис.3.14).

	D	E	F
6	**1-й способ -**		
7	**встроенные ф.**		
8	Среднее		**=СРЗНАЧ(Набор_данных)**
9			
10	Медиана		**=МЕДИАНА(Набор_данных)**
11	Мода		**=МОДА(Набор_данных)**
12	Станд.откл.		**=СТАНДОТКЛОН(Набор_данных)**
13	Дисперсия		**=ДИСП(Набор_данных)**
14	Эксцесс		**=ЭКСЦЕСС(Набор_данных)**
15	Асимметричность		**=СКОС(Набор_данных)**
16	Интервал		
17	Минимум		**=МИН(Набор_данных)**
18	Максимум		**=МАКС(Набор_данных)**
19	Сумма		
20	Объем выборки		**=СЧЁТ(Набор_данных)**

	E	F	G	H	I
49	**Кв**			**Пер**	
50	0	=КВАРТИЛЬ(Набор_данных;E50)		0	=ПЕРСЕНТИЛЬ(Набор_данных;H50)
51	1	=КВАРТИЛЬ(Набор_данных;E51)		0,05	=ПЕРСЕНТИЛЬ(Набор_данных;H51)
52	2	=КВАРТИЛЬ(Набор_данных;E52)		0,5	=ПЕРСЕНТИЛЬ(Набор_данных;H52)
53	3	=КВАРТИЛЬ(Набор_данных;E53)		0,95	=ПЕРСЕНТИЛЬ(Набор_данных;H53)
54	4	=КВАРТИЛЬ(Набор_данных;E54)		1	=ПЕРСЕНТИЛЬ(Набор_данных;H54)

Рис.3.14. Фрагмент рабочего листа MS Excel в режиме отображения формул

При выполнении пункта 5 следует учесть, что функция распределения в данном случае должна быть разрывной. Поскольку наблюдений много (n=60), то ступени мелкие (1/n=1/60), поэтому будем считать функцию распределения непрерывной (при том, что Excel не умеет строить разрывные функции).

Построение графика функции распределения проведем в два этапа:

1) с помощью надстройки MS Excel *«Пакет Анализа - Ранг и персентиль»* получим таблицу с накопленными частотами. Результат работы надстройки приведен на рис.3.15 (интервал ячеек K4:N64);

	K	L	M	N
2	**"Пакет Анализа –**			
3	**Ранг и перцентиль"**			
4	*Точка*	*Столбец1*	*Ранг*	*Процент*
5	60	12,04	1	100,00%
6	59	9,9	2	98,30%
63	2	-0,57	59	1,60%
64	1	-2,18	60	,00%

Рис.3.15 Фрагмент рабочего листа MS Excel с результатами работы надстройки «Пакет Анализа - Ранг и персентиль»

2) полученные значения отсортировать по возрастанию значений вариант в другом диапазоне рабочего листа – в интервале ячеек B132:C194 (рис.3.16);

	B	C
131		**Функция распределения**
132	Значения	Знач.ф.
133	признака	распр.
134	**X**	**F(X)**
135	-2,18	0,00
136	-0,57	0,02
137	-0,36	0,03
138	-0,3	0,05
187	8,23	0,88
188	8,44	0,90
189	9,06	0,92
190	9,09	0,93
191	9,28	0,95
192	9,55	0,97
193	9,9	0,98
194	12,04	1,00
195		

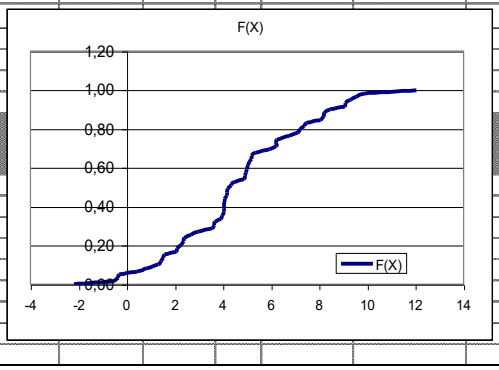

Рис.3.16 Построение функции распределения

Назначить числовой формат данных в ячейках C135:C194, построить график эмпирической функции распределения. Результат представлен на рис. 3.16.

Для выполнения *пункта 6* задания строим диаграмму «Ящик с усами» самостоятельно любым удобным способом (карандаш + линейка + бумага или Excel или средства типа AutoCAD и т.д).

Перед построением заполним таблицу (рис.3.17), предварительно выполнив соответствующие расчеты:

Характеристики "ящика с усами"	
Параметр	Значение
Внешнее верхнее ограждение	
Внутреннее верхнее ограждение	
Верхний ус	
Максимальное значение	
Третья квартиль	
Медиана	
Первая квартиль	
Минимальное значение	
Нижний ус	
Внутреннее нижнее ограждение	
Внешнее нижнее ограждение	
Интерквартильный диапазон	

Рис.3.17. Таблица для построения диаграммы «Ящик с усами»

Для заполнения таблицы следует использовать результаты, полученные в пункте 4, остальные значения (интерквартильный диапазон; внешнее и внутреннее верхнее и нижнее ограждения) вычисляем самостоятельно.

Для выполнения пункта 7 по формулам (3.10 – 3.12) найдем размах, количество и длину интервалов. Результаты вычислений приведены на рис.3.18.

Сам интервальный ряд приведен в ячейках C67:G78 на рис.3.19.

При построении интервального ряда ширина интервала округлена до 2.0, в качестве левой границы первого интервала взято значение –4, что привело к увеличению числа интервалов до 9.

	A	B
67		
68	Размах	
69	14,22	
70		
71	интервалы	
72	кол-во	размер
73	7	2,058814
74		

Рис.3.18 Фрагмент рабочего листа MS Excel. Определение параметров интервального вариационного ряда

При выполнении пункта 8 гистограмму построим двумя способами: с помощью «**Мастера Диаграмм**» и с помощью надстройки Excel «**Данные→Пакет Анализа→Гистограмма»** (рис.3.19).

	C	D	E	F	G	H	I	J	K	L
66	**Построение интервального ряда**									
67		Границы инетрвала		середина интервала	частота	Накопленная частота	Относительная частота	Накопленная отн.частота	Среднее	Дисперсия
68		левая	правая							
69	№	a[i]	b[i]	x[i]	mi		mi/N		mi*xi	mi*(xi-x ср)^2
70	1	-4	-2	-3,00	1	1	0,0167	0,0167	-3	60,32
71	2	-2	0	-1,00	3	4	0,0500	0,0667	-3	99,76
72	3	0	2	1,00	6	10	0,1000	0,1667	6	85,13
73	4	2	4	3,00	12	22	0,2000	0,3667	36	37,45
74	5	4	6	5,00	20	42	0,3333	0,7000	100	1,09
75	6	6	8	7,00	8	50	0,1333	0,8333	56	39,90
76	7	8	10	9,00	9	59	0,1500	0,9833	81	161,29
77	8	10	12	11,00	0	59	0,0000	0,9833	0	0,00
78	9	12	14	13,00	1	60	0,0167	1,0000	13	67,79
79	sum				60				286	552,73
80	average								4,77	9,21
81	Станд.откл.									3,035164283

Рис.3.19. Фрагмент рабочего листа MS Excel. Построение интервального вариационного ряда

Обратим внимание, что это средство позволяет получить значения частот, не обращаясь к их непосредственному подсчету.

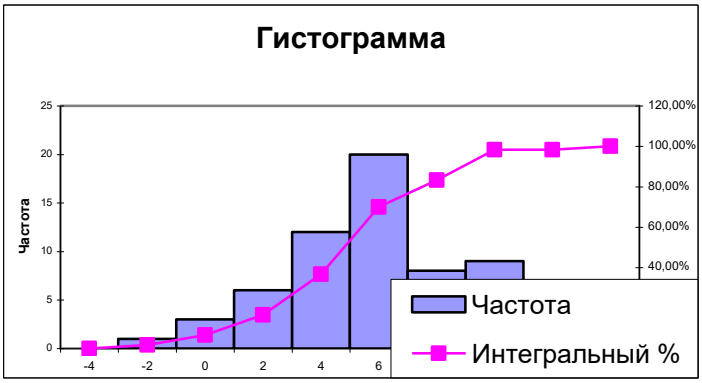

Рис.3.20 Диаграмма MS Excel, полученная с помощью надстройки «Пакет Анализа - Гистограмма»

При выполнении пункта 9 для вычисления значений m_X, D_X, σ_X заполним интервал ячеек K70:L78. Результаты вычислений представлены в ячейках K80, L80-L81.

СПИСОК ЛИТЕРАТУРЫ

1. *Гмурман В.Е.* Теория вероятностей и математическая статистика, изд.12. - М.: Юрайт, 2020, с.479.
2. *Ивченко Г.И.* Математическая статистика /Г.И.Ивченко, Ю.И.Медведев. Учебник. – М.:Книжный дом»ЛИБРОКОН», 2014. – 352 с.
3. *Бер К., Кэйри П.* Анализ данных с помощью Microsoft Excel. - М.: Вильямс, 2005, с. 560.

ОГЛАВЛЕНИЕ