ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Цель : изучение методики предварительной обработки экспериментальных данных, проверки соответствия распределения результатов измерения закону нормального распределения; изучение возможностей пакета MS Excel при решении задач статистической обработки экспериментальных данных.

Общие положения

Объект исследования – это объект любого характера, который изучается экспериментальным путем.

Эксперимент – это специальным образом спланированная и организованная процедура изучения некоторого объекта исследования, при которой на этот объект оказывают запланированные воздействия и регистрируют его реакции на эти воздействия.

Экспериментальные данные – все исходные и выходные числовые данные эксперимента, сведенные в таблицу экспериментальных данных.

Обработка экспериментальных данных – различные методы построения математической модели объекта по таблице экспериментальных данных.

Основным «**рабочим инструментом**» эксперимента и обработки экспериментальных данных является численное значение факторов воздействия и откликов объекта исследования, т. е. **число**.

Числа при экспериментировании получают тремя способами:

- **подсчетом;**
- **измерением;**
- **методом экспертных оценок**.

Предварительная обработка результатов измерений и наблюдений необходима для того, чтобы в дальнейшем, при построении эмпирических зависимостей, **эффективно использовать статистические методы и корректно анализировать полученные результаты**.

Содержание предварительной обработки в основном состоит в **отсеивании грубых погрешностей** измерения или погрешностей, неизбежно имеющих место при переписывании цифрового материала или при вводе на электронный носитель информации.

Другим важным моментом предварительной обработки данных является **проверка соответствия** распределения результатов измерения **закону нормального распределения**.

Если эта гипотеза неприемлема, то следует определить, какому закону распределения подчиняются опытные данные, и, если это возможно, **преобразовать данное распределение к нормальному**.

Только после выполнения перечисленных выше операций можно перейти **к построению эмпирических формул**, применяя, например, **метод наименьших квадратов**.

Генеральная совокупность и выборка.

Генеральной называют совокупность всех мыслимых наблюдений, которые могли бы быть сделаны при данном комплексе условий.

Генеральная совокупность может быть **конечной и бесконечной**.

Данное выше определение генеральной совокупности можно считать строго обоснованным только для случаев конечных генеральных совокупностей

Понятие **бесконечной генеральной** совокупности – математическая **абстракция**, как и представление о том, что измерить случайную величину можно бесконечное число раз.

Приближенно бесконечную генеральную совокупность можно истолковать как **предельный случай** конечной генеральной совокупности.

В распоряжении исследователя, **никогда нет генеральной совокупности**, он может изучать только ее часть – выборку, причем всегда ограниченного объема.

Результаты ограниченного ряда наблюдений x_1, x_2, ..., x_n случайной величины можно рассматривать как в**ыборку** из данной генеральной совокупности.

Выборка – любое конечное подмножество генеральной совокупности, предназначенное для непосредственных исследований,

Объем – количество единиц в выборке.

Относительной частотой случайного события, называется отношение числа появлений этого события к общему числу произведенных испытаний.

Мера объективной возможности случайного события называется вероятностью случайных событий.

Относительные частоты можно **истолковать как выборочные значения** вероятностей случайных событий.

Характеристики **теоретических распределений** можно рассматривать как характеристики, существующие в **генеральной совокупности**, а характеристики **эмпирических распределений** – как **выборочные характеристики.**

Можно встретить и другую терминологию. Характеристики распределения вероятностей в генеральной совокупности называют п**араметрами**, а выборочные (эмпирические) значения характеристик – **оценками или статистиками.**

Параметры обозначаются буквами греческого алфавита, а оценки – соответствующими буквами латинского алфавита.

Исходными данными при оценивании, как и при проверке любых предположений (статистических гипотез), касающихся неизвестного распределения случайной величины могут быть лишь только те результаты наблюдений, которые были получены **в ходе проведения опытов** (на выборке ограниченного объема).

Причем **предварительная обработка экспериментальных** данных обычно **начинается с подсчета тех или иных функций от результатов наблюдений (статистик).**

Оценивание – определение приближенного значения неизвестного параметра генеральной совокупности по результатам наблюдений.

К оценкам предъявляются требования состоятельности, **несмещенности, эффективности.**

Состоятельная оценка – оценка, сходящаяся по вероятности к значению оцениваемого параметра при безграничном возрастании объема выборки.

Несмещенная оценка – оценка, **математическое ожидание** которой равно значению оцениваемого параметра.

Оценка параметра называется э**ффективной**, если среди прочих оценок того же параметра она обладает **наименьшей дисперсией**.

Вычисление характеристик эмпирических распределений (выборочных характеристик).

Здесь и в дальнейшем речь идет только о непрерывно распределенных случайных величинах.

Пусть имеется ограниченный ряд наблюдений x_1, x_2, ..., x_n случайной величины. Среднее значение наблюдаемого признака определяется по формуле

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

где n – количество x_i значений выборки или объем выборки;
x_i - результат измерения i-й единицы.

Таким образом, $\bar{x}$ представляет собой эмпирическое или выборочное среднее. Если вычислено среднее, то легко найти отклонение каждого наблюдения от среднего

$$d_i = x_i - \bar{x}.$$

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

называют дисперсией или вторым центральным моментом эмпирического распределения $S^2 = m_2$.

В случае одномерного эмпирического распределения произвольным моментом порядка k называется сумма k-х степеней отклонений результатов наблюдений от произвольного числа с, деленная на объем выборки п:

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

где k может принимать любые значения натурального ряда чисел.

Первый центральный момент

$$m_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$$

Второй центральный момент $m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Несмещенную оценку для S^2 (или σ^2 - дисперсия теоретического распределения) можно найти по формуле

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

Выборочные среднеквадратические отклонения соответственно могут быть найдены по формулам

$$\overline{S} = \sqrt{\overline{S}^2}; S = \sqrt{S^2}.$$

Из других моментов чаще всего используют моменты третьего и четвертого порядка:

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \quad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Выборочное значение коэффициента вариации V, являющееся мерой относительной изменчивости наблюдаемой случайной величины в %, определяют по формуле

$$V = \frac{\overline{S}}{\bar{x}} \cdot 100\%.$$

Для нормальных и близких к нормальному распределений показатель V служит индикатором однородности выборочных наблюдений: принято считать, что при выполнимости неравенства V ≤ 33 % выборка является **количественно однородной по данному признаку**.

Выборочные значения характеристик распределения имеет смысл вычислять только в случае, если выборка является случайной.

Обычно на практике наблюдаемые значения x_1, x_2, ..., x_n величины случайные и отклонения их от среднего значения обусловлены погрешностями измерения и т. д. В свою очередь, погрешности – результат действия многих факторов.

Если имеет место такой редкий случай, когда в распоряжении исследователя имеется вся генеральная совокупность и необходимо сделать из нее выборку, то используют один из методов рандомизации (случайного выбора).

Отсев грубых погрешностей.

Можно встретить большое количество различных рекомендаций для проведения отсева грубых погрешностей наблюдения (аномальных значений).

Рассмотрим **наиболее простой метод отсева грубых погрешностей**. Если в распоряжении экспериментатора имеется выборка небольшого объема, то можно воспользоваться методом вычисления максимального относительного отклонения:

$$\tau = \frac{\left| x_{\min(\max)} - \bar{x} \right|}{\bar{S}} \leq \tau_{1-\alpha},$$

где $x_{min(max)}$ - крайний (наибольший или наименьший) элемент выборки, по которой подчитывается $\bar{x}$, S и τ, вычисленной при доверительной вероятности $p = 1 - \alpha$.

Таким образом, для выделения аномального значения вычисляют τ, которое затем сравнивают с табличным значением $\tau_{1-\alpha}$.

Если это неравенство $\tau < \tau_{1-\alpha}$ соблюдается, то наблюдение не отсеивают, если не соблюдаются, то наблюдение исключают.

После исключения того или иного наблюдения или нескольких наблюдений характеристики эмпирического распределения должны быть пересчитаны по данным сокращенной выборки.

Квантили распределения статистики τ при уровнях значимости $\alpha = 0{,}10$, $\alpha = 0{,}05$, $\alpha = 0{,}025$, $\alpha = 0{,}01$ или доверительной вероятности $p = 1 - \alpha = 0{,}90$; 0,95; 0,975; 0,99 даны в справочниках.

На практике обычно используют уровень значимости $\alpha = 0{,}05$ (результат получается с 95%-й доверительной вероятностью).

Процедуру отсева нужно повторить и для следующего по абсолютной величине максимального относительного отклонения, но предварительно необходимо пересчитать x и S для выборки нового объема $(n-1)$.

Полигон и гистограмма частот распределения.

Если полученные экспериментальные данные разделить на классы, то можно построить полигон и гистограмму частот.

Разбиение на классы можно выполнить по правилу Штюргеса с округлением полученного значения до ближайшего целого числа.

Число классов определяется по формуле

Jock Sturges

$$k \approx 1 + 3{,}32 \cdot \lg(n).$$

Далее определяют размах варьирования:

$$R = x_{\max} - x_{\min}$$

Определяют ширину интервала $h = \frac{R}{k}.$

Затем устанавливают границы интервалов и подсчитывают число попаданий случайной величины в каждый из выбранных интервалов (абсолютные частоты Bj), для этого значения экспериментальных данных просматривают по порядку от первой до последней строчки, и при чтении каждого результата соответствующую метку (точку или черточку) заносят в тот класс, к которому относится данное наблюдение. Каждая метка соответствует одному значению из выборки.

Затем определяют относительные частоты попаданий в j-й интервал (класс) как (B_j / n) и относительные накопленные частоты как $\Sigma(B_j / n)$.

Для проверки, сумма B_j равна количеству экспериментальных данных (опытов) n.

Гистограмма и полигон распределений являются графическим отображением частот, которые, в свою очередь, представляют собой оценки плотностей вероятностей.

Кумулятивная линия – график накопленных частот, в свою очередь оценивающих функцию распределения F(x) в точке x. Многие наблюдения в природе при такой обработке дают колоколообразные полигоны распределения.

Если распределение случайной величины подчиняется определенному закону и может быть хотя бы **приближенно описано кривой** $y = ae^{-bx2}$, то такое распределение называют нормальным.

Так как к коэффициентам a и b предъявляется только одно требование, а именно: a, b > 0, то можно говорить о семействе кривых нормального распределения.
С увеличением коэффициента a кривая «вытягивается» в высоту; при увеличении коэффициента b кривая «сплющивается».

Нормальное распределение обладает и другими важными свойствами, которые позволяют считать это распределение основой математической статистики. Рассмотрим эти свойства.

1. Ордината y, которая определяет высоту кривой для каждой точки оси Ox (абсциссы), представляет собой плотность вероятности некоторого значения переменной x и определяется следующей формулой

$$y = f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)}{2\sigma}}$$

$$-\infty < x < +\infty, \sigma > 0,$$

где σ – среднеквадратическое отклонение теоретического распределения;
μ – среднее значение (математическое ожидание) теоретического распределения.

Из формулы (16) следует, что нормальное распределение полностью определяется величинами μ и σ
(π = 3,141593... и e = = 2,718282... – математические постоянные).

Математическое ожидание μ определяет положение кривой распределения относительно оси Ox.

Среднеквадратическое отклонение σ определяет форму кривой.

Чем больше σ (разброс данных), тем кривая становится более пологой (ее основание более широкое).

2. Кривая нормального распределения симметрична относительно среднего значения.

3. Максимум ординаты кривой $$y_{\max} = \frac{1}{\sqrt{2\pi\sigma^2}}$$

что при σ = 1 составляет примерно 0,4. Если x → ± ∞, то y →0 (асимптотически).

Другими словами, очень большие и очень малые значения переменной x маловероятны.

Примерно 2/3 всех наблюдений лежит в площади, отсекаемой перпендикулярами к оси Ox ($\mu \pm \sigma$).

При большом объеме выборки примерно 90 % всех наблюдений лежит между $-1{,}64\sigma$ и $+1{,}64\sigma$. Границы $-0{,}675\sigma$ и $+0{,}675\sigma$ называют вероятными отклонениями: в этом интервале находится около 50 % всех наблюдений.

Для нормального распределения среднее, мода и медиана совпадают.

Медианой выборки является среднее значение из всего упорядоченного набора значений.

Модой выборки называется значение, которое встречается большее число раз в выборке.

Для статистических методов построения эмпирических зависимостей очень важно, чтобы результаты наблюдений подчинялись нормальному закону распределения, поэтому проверка нормальности распределения – основное содержание предварительной обработки результатов наблюдений.

Проверка гипотезы нормальности распределения.

.Среднее абсолютное отклонение.

Для небольших выборок (n < 120) можно найти простые рекомендации по проверке нормальности распределения.

Для этого необходимо вычислить **среднее абсолютное отклонение (САО)** по формуле

$$CAO = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Для выборки, имеющей приближенно нормальный закон распределения, должно быть справедливо выражение

$$\left|\frac{CAO}{\bar{S}} - 0{,}7979\right| < \frac{0{,}4}{\sqrt{n}}.$$

Пользуясь САО, можно также с 95%-й доверительной вероятностью оценить μ (среднее значение теоретического распределения) по $\overline{x}$:

$$\mu = \overline{x} \pm (0{,}71 \div 0{,}6) \cdot CAO.$$

Коэффициент (0,71 ÷ 0,6) зависит от величины выборки n (в данном случае n = 15÷ 20) и 1 - α = 0,95.

Коэффициенты для определения 95%-х доверительных границ для среднего значения по САО приведены в справочниках.

Размах варьирования R.

Быструю проверку гипотезы нормальности распределения для сравнительно широкого класса выборок 3 < n < 1000 можно выполнить, используя размах варьирования R.

Подсчитываем отношение $\frac{R}{\overline{S}}$

и сопоставляем с критическими верхними и нижними границами этого отношения, приведенными в справочниках

Если $k_{i'} \le \frac{R}{\overline{S}} \le k_{\hat{a}}.$

меньше нижней или больше верхней границы, то нормального распределения нет. Особенно важно, чтобы это условие соблюдалось при α = 0,10 (10%-й уровень значимости).

Показатели асимметрии и эксцесса.

Некоторое представление о близости эмпирического распределения к нормальному может дать анализ показателей асимметрии и эксцесса. Показатель асимметрии можно определяют по формуле

$$g_1 = \frac{m_3}{m_2^{3/2}}.$$

Для симметричных распределений $m_3 = 0$ и $g_1 = 0$.

Для нормального распределения $m_4 / m_2^2 = 3$.

Для удобства сравнения эмпирического распределения и нормального в качестве показателя эксцесса принимают величину

$$g_2 = \frac{m_4}{m_2^2} - 3.$$

Несмещенные оценки для показателей асимметрии G_1 и эксцесса G_2 определяют соответственно по формулам:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1;$$

$$G_2 = \frac{n-1}{(n-2)(n-3)}\left[(n+1)g_2 + 6\right].$$

Для проверки гипотезы нормальности распределения следует также вычислить среднеквадратические отклонения для показателей асимметрии и эксцесса соответственно:

$$S_{G_1} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}};$$

$$S_{G_2} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}.$$

Если выполняются условия $G_1 \le 3S_{G1}$, $G_2 \le 5S_{G2}$, то гипотеза нормальности исследуемого распределения может быть принята.

По критерию χ^2 (хи-квадрат)

Рассмотрим методику проверки гипотезы нормальности распределения по χ^2 критерию. Применение критерия χ^2 предполагает также использование свойств так называемого стандартного нормального распределения, которое имеет вид:

$$y = f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \cong 0{,}4e^{\frac{-z^2}{2}}.$$

Расчеты выполняют в табличной форме. Значения χ^2 определяют по формуле

$$\chi^2 = \sum_{j=1}^{n} \frac{(B_j - E_j)^2}{E_j},$$

где B_j – наблюдаемая частота;
E_j – ожидаемая по стандартному нормальному распределению частота.

$$E_j = f(z_j) \cdot k',$$

$$k' = \frac{nn_{\bar{s}}}{\bar{S}},$$

где f(zj) – уравнение кривой стандартного нормального распределения:

$$f(z_j) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_j^2}{2}};$$

z_j – степень функции кривой нормального распределения:

$$z_j = \frac{\left| x_i - \bar{x}_{ожид} \right|}{\bar{S}_{ожид}}$$

$\bar{x}_{ожид}$ ожидаемое среднее значение наблюдаемого признака

$$\bar{x}_{ожид} = \frac{1}{n}\sum_{j=1}^{n_{кл}} B_j x_j;$$

$\bar{S}_{ожид}$ ожидаемая дисперсия:

$$\bar{S}_{ожид}\sqrt{\frac{\sum_{j=1}^{n_{кл}} B_j x_j^2 - \frac{(\sum_{j=1}^{n_{кл}} B_j x_j)^2}{n}}{n-1}};$$

$n_{кл}$ - число классов (интервалов).

Полученное значение χ^2 сравнивают с табличным или критическим значением $\chi^2_{\text{пкα}}$,.

Число степеней свободы v определяют по формуле

$$\nu = n_{кл} - 1 - \kappa,$$

где r – число параметров распределения (для нормального распределения r = 2, так как оцениваются два параметра $\overline{x}, \overline{S}$

Гипотеза нормальности распределения принимается в случае выполнения условия $\chi^2 \leq \chi^2_{\text{пкα}}$.

Методика проверки нормальности распределения по показателям асимметрии и эксцесса очень хорошо иллюстрирует использование моментов, а также удобна при использовании компьютерных технологий.

Для практического применения (особенно при расчетах с использованием компьютерных технологий) рекомендуются в основном две методики: по размаху варьирования и по χ^2-критерию, причем первая служит для быстрой «прикидочной» проверки, а вторая – для основательной проверки нормальности распределения.

В настоящее время обработку экспериментальных данных существенно облегчают современные компьютерные технологии, современное программное обеспечение. Например, электронные таблицы MS Excel.

Особенности использования средств инструмента «Описательная статистика» в надстройке «Пакет анализа» MS Excel

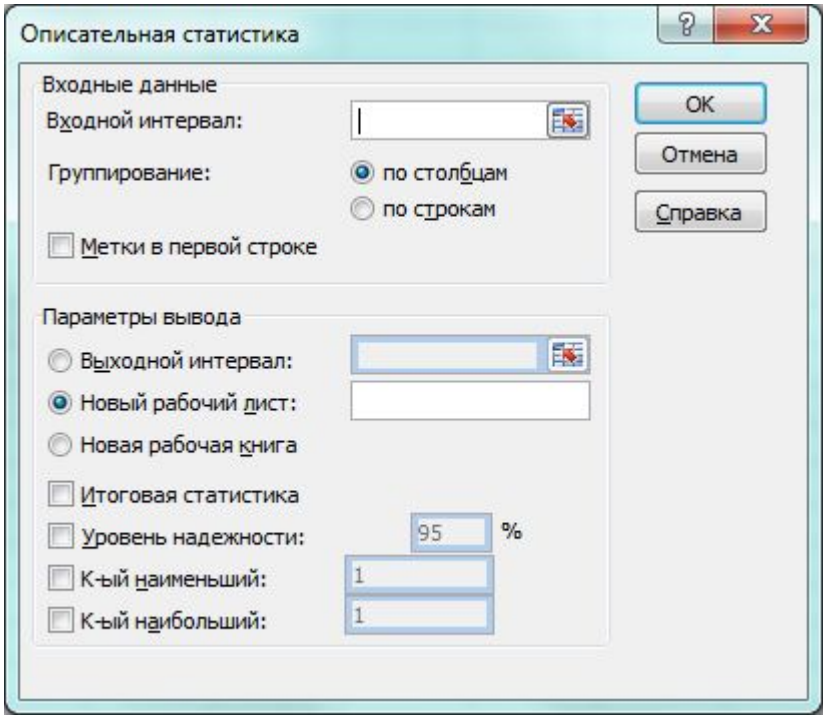
В состав MS Excel входит надстройка «Пакет анализа», которая содержит 19 статистических процедур и около 50 функций.

Для анализа данных с помощью средств этого пакета следует указать входные данные и выбрать параметры; в итоге расчет будет выполнен с помощью подходящей статистической или инженерной макрофункции, а результат будет помещен в выходной диапазон.

Для доступа к этим инструментам необходимо в меню «Данные» нажать кнопку **«Анализ данных**». Если кнопка «Анализ данных» недоступна, необходимо загрузить надстройку «Пакет анализа».

Инструмент «Описательная статистика» (вместе с инструментом «Гистограмма», алгоритм использования которого будет описан далее) является наиболее часто используемым из всего «Пакета анализа», поскольку быстро и просто вычисляет основные статистические характеристики одномерных выборок

В «Пакете анализа» инструмент «Описательная статистика» используется для генерации одномерного статистического отчета, который включает ряд показателей положения, вариации и формы распределения признаков выборочной и генеральной совокупностей, а также среднюю и предельную ошибки выборки для средней.

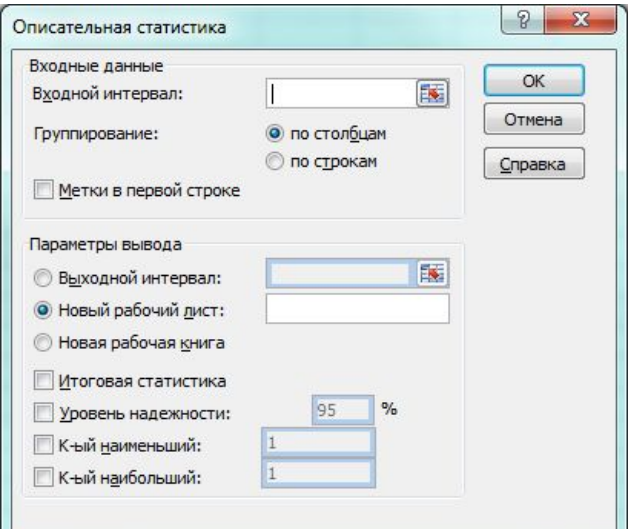


После выбора инструмента «Описательная статистика» в по-явившемся диалоговом окне инструмента (рис. задаются следующие параметры:

.Поле Входной интервал – вводится ссылка на диапазон ячеек, содержащих значения анализируемого признака или признаков.

В качестве входного интервала может быть указан диапазон, который содержит ряды значений сразу нескольких анализируемых признаков. В таком случае показатели «Описательной статистики» будут рассчитаны для каждого ряда и представлены в единой таблице в виде отдельных столбцов.

2. Переключатель Группирование: по столбцам/строкам – устанавливается в положение по столбцам или по строкам в зависимости от того, в каком направлении располагаются анализируемые данные во входном диапазоне – вертикальном (по столбцам) или горизонтальном (по строкам).



3. Флажок Метки в первой строке – устанавливается в активное состояние, если первая строка во входном диапазоне содержит заголовки.

Если заголовки отсутствуют, поле не активизируется. В этом случае будут автоматически созданы стандартные названия для данных выходного диапазона.

4. Поле Выходной интервал – вводится ссылка на ячейку заголовка первого столбца выходной результативной таблицы.

Размер выходного диапазона ячеек определяется автоматически. В случае возможного наложения выходного диапазона на другие данные на экране появится соответствующее сообщение.

5. Переключатели Новый рабочий лист и Новая рабочая книга – устанавливаются в активное положение при необходимости открытия соответственно нового листа или новой книги.

В новом листе результаты анализа располагаются начиная с ячейки А1, в новой книге – на первом листе начиная с ячейки А1.

6. Флажок Итоговая статистика – устанавливается в активное состояние, если для данных входного диапазона необходимо произвести расчет основных показателей

7. Флажок Уровень надежности – устанавливается в активное состояние, если в результативную таблицу необходимо включить строку для оценки предельной ошибки выборки с заданной доверительной вероятностью.

Значение уровня надежности выражается в процентах и задается в поле напротив флажка Уровень надежности. Уровень надежности 95,0 % (что равносильно доверительной вероятности p = 0,95 или же уровню значимости α = 0,05) фиксируется в поле автоматически.

8. Флажки K-й наименьший и K-й наибольший – активизируются, если в результативную таблицу необходимо включить строку соответственно для k-го наименьшего (начиная с минимума x_{min}) и k-го наибольшего (начиная с максимума x_{max}) значений элементов в выборке.

В этом случае в поле, расположенном напротив каждого флажка, вводится число k. При k = 1 выходные строки будут содерать соответственно x_{min} и x_{max}.

Между терминологией инструмента «Описательная статистика» и терминами, принятыми в отечественной статистике, имеется ряд расхождений.

Вычисленные значения всех вышеперечисленных показателей Excel. При этом показатели могут рассчитываться сразу для нескольких рядов данных в соответствии с заданным входным диапазоном ячеек.

Следует обратить внимание на то, что расчет параметров в режиме «Описательная статистика» имеет ряд важных особенностей:

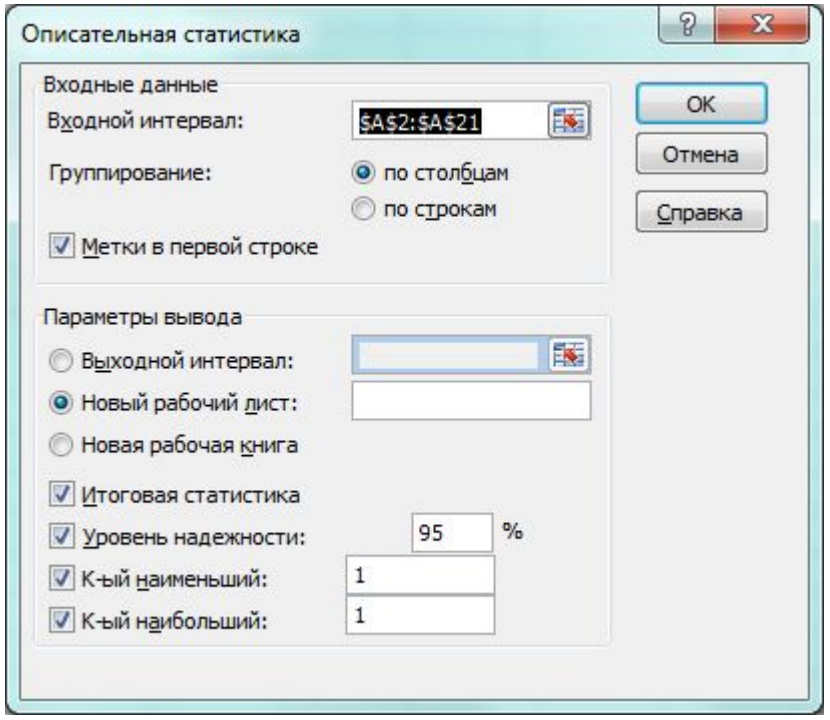
1. В качестве значений параметров: Стандартное отклонение, Дисперсия выборки, Эксцесс, Асимметричность – MS Excel генерирует оценки соответствующих параметров для генеральной совокупности, а не для выборки.

2. Для применения «Описательной статистики» предварительное ранжирование исходных данных не требуется: при вычислении показателей ранжирование выполняется автоматически.

3. Появление в ячейке Мода индикатора ошибки #Н/Д указывает на то, что в анализируемых данных нет одинаковых значений признака. В этом случае в качестве моды Мо выбирается то значение признака, которое соответствует максимальной ординате теоретической кривой распределения.

4. Индикатор ошибки # ДЕЛ/0! в ячейке Эксцесс и/или Асимметричность означает, что в результативной таблице стандартное отклонение является нулевым или же заданный входной диапазон данных содержит менее четырех элементов данных

44
32
45
38
50
48
33
42
33
44
53
42
37
52
52
39
40
39
44
39



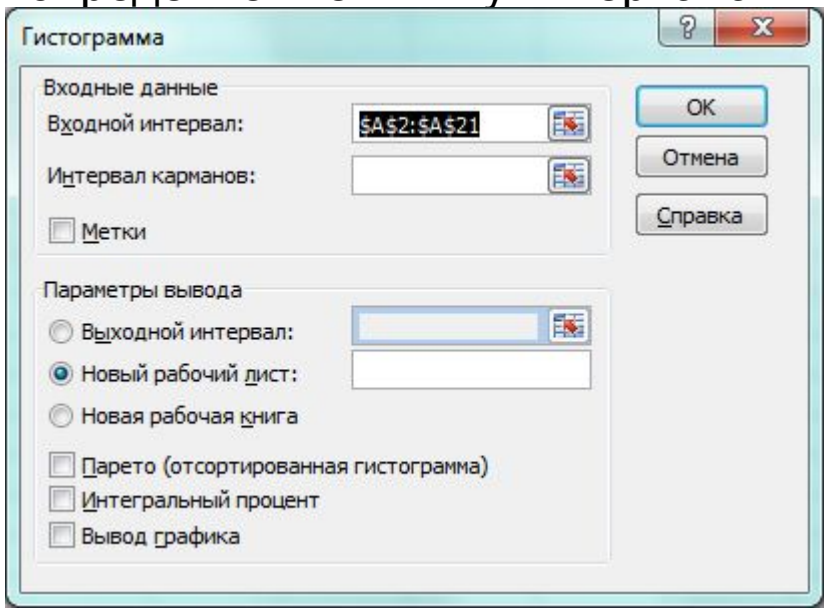
	A	B
1	44	
2		
3	Среднее	42,21053
4	Стандартная ошибка	1,501615
5	Медиана	42
6	Мода	39
7	Стандартное отклонение	6,545388
8	Дисперсия выборки	42,84211
9	Эксцесс	-0,88991
10	Асимметричность	0,194583
11	Интервал	21
12	Минимум	32
13	Максимум	53
14	Сумма	802
15	Счет	19
16	Наибольший(1)	53
17	Наименьший(1)	32
18	Уровень надежности(95,0%)	3,154776
19		

Особенности использования средств инструмента «Гистограмма» в надстройке «Пакет анализа» MS Excel

В надстройке Excel «Пакет анализа» инструмент «Гистограмма» используется для генерации интервального вариационного ряда с равными по величине интервалами, а также для построения гистограммы и кумуляты сформированного ряда распределения.
Инструмент «Гистограмма» производит следующие действия:

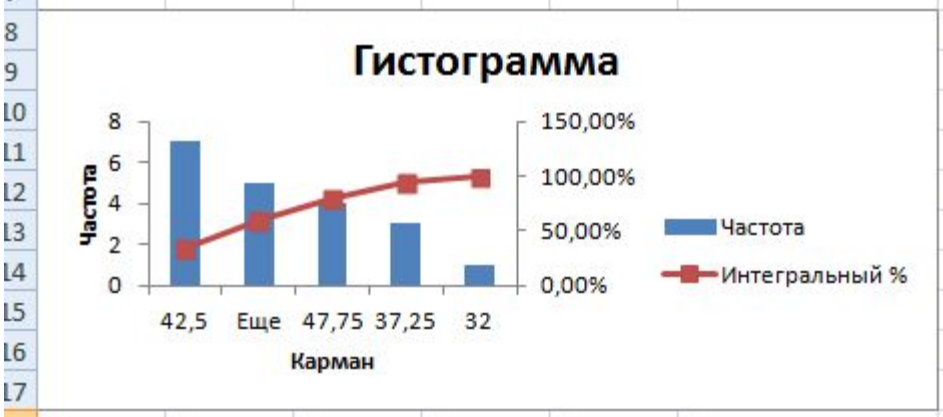
- рассчитывает число интервалов по формуле ; $k \approx 1 + 3{,}32 \cdot \lg(n).$

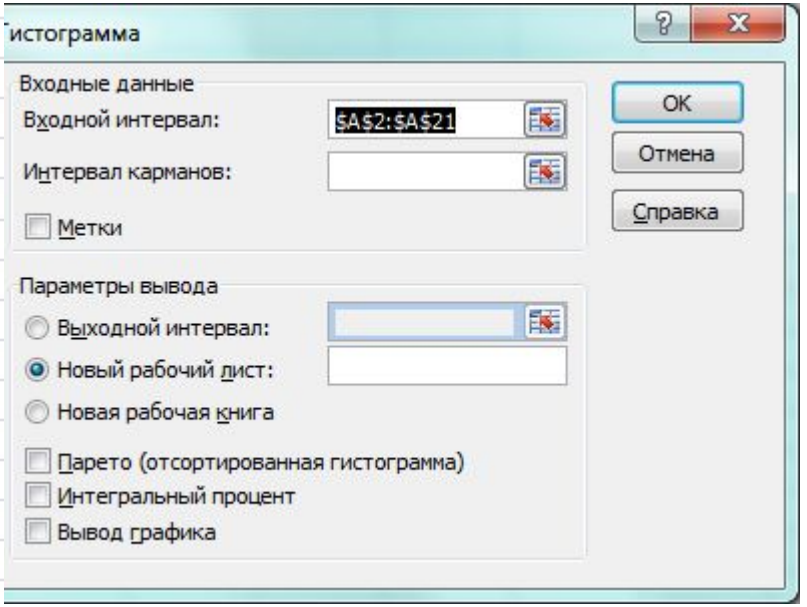
- определяет величину интервала h по формуле $h = \frac{R}{k-1}$



- определяет нижние границы интервалов;
- формирует интервальный вариационный ряд в соответствии с
- величинами k, h;
- рассчитывает частоты и накопленные частоты интервалов,
- определяя число попаданий данных в сформированные интервалы;
- строит столбиковую диаграмму частот (которая может быть
- преобразована в гистограмму) и кумуляту накопленных частот для
- полученного ряда распределения;
- генерирует для вариационного ряда выходную таблицу.

	Карман	Частота	Интегральный	Карман	Частота	Интегральный %
2	32	1	5,00%	42,5	7	35,00%
3	37,25	3	20,00%	Еще	5	60,00%
4	42,5	7	55,00%	47,75	4	80,00%
5	47,75	4	75,00%	37,25	3	95,00%
6	Еще	5	100,00%	32	1	100,00%





статистическая интерпретация терминологии инструмента «Гистограмма»

Термин инструмента «Гистограмма»	Термин, принятый в статистике
Карманы	Интервалы вариационного ряда
Интервал карманов	Диапазон ячеек, содержащих в возрастающем порядке верхние границы интервалов
Интегральный процент	Накопленная частота, выраженная в процентах

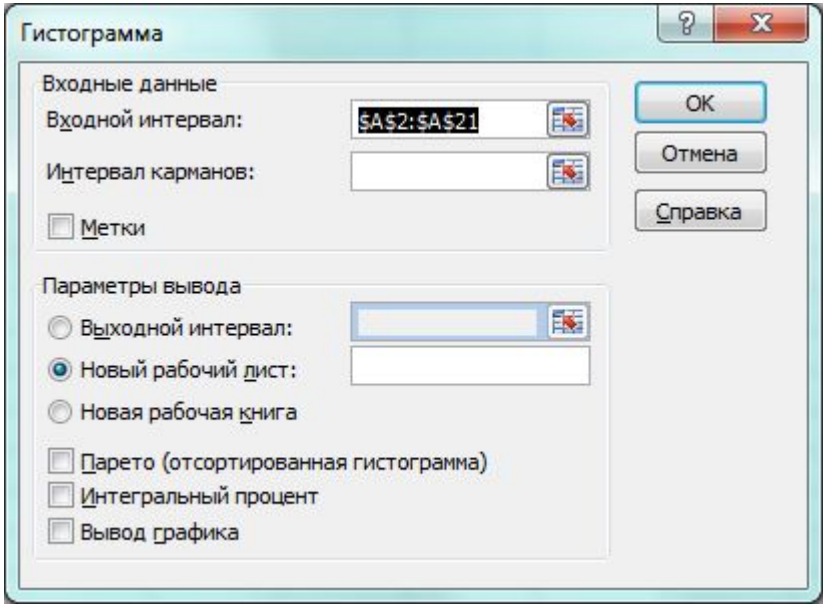
Инструмент «Гистограмма» имеет два режима работы:

- режим автоматического формирования интервалов вариационного ряда, имеющих равную величину h;

- режим формирования интервалов ряда в соответствии с границами, заданными пользователем.

Если при этом заданные интервалы будут не равны между собой, то в сгенерированной столбиковой диаграмме частоты попадания в интервал не будут связаны с размером интервала, что не позволит правильно оценить характер распределения единиц изучаемой совокупности.

Запуск инструмента «Гистограмма» осуществляется аналогично инструменту «Описательная статистика» надстройки «Пакет анализа».
В появившемся диалоговом окне инструмента «Гистограмм задаются следующие параметры:



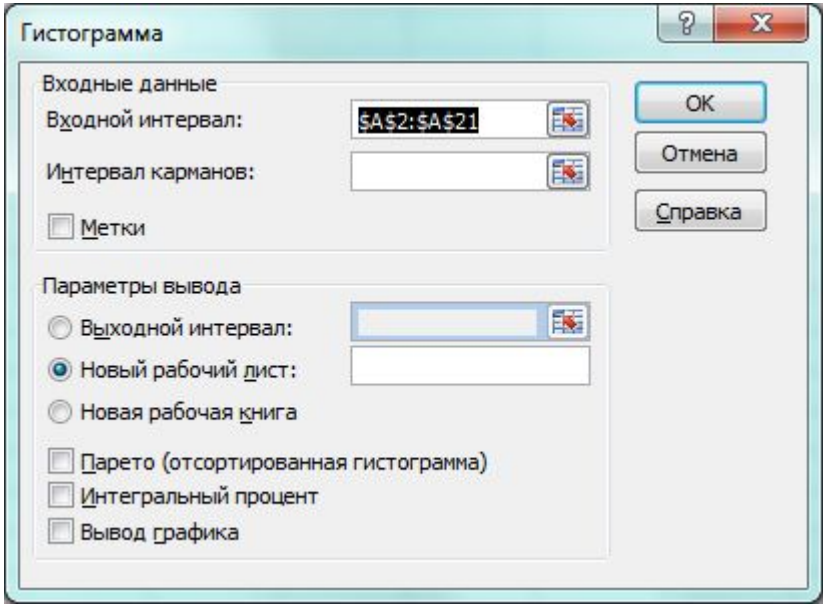
1. Поле Входной интервал – вводится ссылка на диапазон ячеек, содержащих значения анализируемого признака.
2. Интервал карманов (необязательный параметр) – вводится ссылка на диапазон ячеек, в которых задаются верхние границы интервалов. Если такой диапазон не указан, Excel осуществляет расчет нижних границ интервалов автоматически.
3. Флажок Метки не активизируется.
4. Поле Выходной интервал – вводится ссылка на ячейку заголовка первого столбца формируемой таблицы интервального вариационного ряда.

5. Переключатель Новый рабочий лист/Новая рабочая книга – открывает Новый рабочий лист/Новую рабочую книгу.

6. Флажок Парето (отсортированная гистограмма) – устанавливается в активное состояние при необходимости представить данные в порядке убывания частоты. Если флажок снят, то данные в выходном диапазоне будут приведены в порядке следования интервалов.

7. Флажок Интегральный процент – устанавливается в активное состояние, если необходимо рассчитать накопленные частоты (выраженные в процентах) и построить график кумуляты.

8. Флажок Вывод графика – устанавливается в активное состояние при необходимости автоматического построения столбиковой диаграммы.



Необходимо отметить, что инструменты «Пакет анализа» имеют определенные ограничения и иногда удобнее воспользоваться статистическими функциями или другими средствами MS Excel.

Преимуществом функций перед данными средствами является то, что функции автоматически пересчитываются при любых изменениях, сделанных в выборке, тогда как эти средства необходимо выполнять заново, если выборка изменилась.

пример выполнения

Исходные данные демонстрационного примера:
Данные наблюдения роста группы двадцатилетних юношей-студентов-третьекурсников (табл. 3).

1. Определяем среднее значение выборки
2. Определяем дисперсию
3. Среднеквадратичное отклонение

	A	B	C	D
1	№ п/п	X;, см		
2	1	183		
3	2	170		
4	3	176		
5	4	178		
6	5	176		
7	6	180		
8	7	176		
9	8	185		
10	9	184		
11	10	174		
12	11	168		
13	12	174		
14	13	189		
15	14	172		
16	15	175		
17	16	167		
18	17	179		
19	18	276		
20	19	169		
21	20	178		
22	X	3629		
23	x сред	181,45	СРЗНАЧ(B2:B21)	
24	дисперсия	529,3132	ДИСП(B2:B21)	
25	ср.кв.отклонение	23,00681	КОРЕНЬ(B24)	
26	xmin	167	МИН(B2:B21)	
27	xmax	276	МАКС(B2:B21)	
28	$\lvert x_{\min} - \bar{x} \rvert$	14,45	ABS(B26-B23)	
29	$\lvert x_{\max} - \bar{x} \rvert$	94,55	ABS(B27-B23)	
30	Исключить ошибку	4,109653	МАКС(B28:B29)/B25	τ1−α = 2,62;

4.Определяем отсев грубых погрешностей.

Для отсева погрешностей используем относительного отклонения.

Условие отсева

$$\tau = \frac{\left|x_{\min(\max)} - \bar{x}\right|}{\bar{S}} \le \tau_{1-\alpha}$$

Выбираем наибольший

$$\left|x_{\min(\max)} - \bar{x}\right|$$

$$\sum_{\text{mod}} = \sum_{i=1}^{n} \left|x - \bar{x}\right|$$

	A	B	C	E	F	G	H
1	№ п/п	X;, см	$x - \bar{x}$	C2^2	C2^3	C2^4	
2	1	183	6,526315789	42,5928	277,974	1814,146	
3	2	170	-6,473684211	41,90859	-271,303	1756,33	
4	3	176	-0,473684211	0,224377	-0,10628	0,050345	
5	4	178	1,526315789	2,32964	3,555766	5,427222	
6	5	176	-0,473684211	0,224377	-0,10628	0,050345	
7	6	180	3,526315789	12,4349	43,84939	154,6268	
8	7	176	-0,473684211	0,224377	-0,10628	0,050345	
9	8	185	8,526315789	72,69806	619,8466	5285,008	
10	9	184	7,526315789	56,64543	426,3314	3208,705	
11	10	174	-2,473684211	6,119114	-15,1368	37,44355	
12	11	168	-8,473684211	71,80332	-608,439	5155,717	
13	12	174	-2,473684211	6,119114	-15,1368	37,44355	
14	13	189	12,52631579	156,9086	1965,487	24620,3	
15	14	172	-4,473684211	20,01385	-89,5356	400,5542	
16	15	175	-1,473684211	2,171745	-3,20047	4,716477	
17	16	167	-9,473684211	89,75069	-850,27	8055,187	
18	17	179	2,526315789	6,382271	16,12363	40,73339	
19	19	169	-7,473684211	55,85596	-417,45	3119,888	
20	20	178	1,526315789	2,32964	3,555766	5,427222	
21	X	3353					
22	∑mod		88,42105263				
23	x сред	176,4737	СРЗНАЧ(B2:B21)				
24	дисперсия	35,92982	ДИСП(B2:B21)				
25	ср.кв.отклонение	5,994149	КОРЕНЬ(B24)				
26	xmin	167	МИН(B2:B21)				
27	xmax	189	МАКС(B2:B21)				
28	$\left\|x_{\min} - \bar{x}\right\|$	9,473684	ABS(B26-B23)				
29	$\left\|x_{\max} - \bar{x}\right\|$	12,52632	ABS(B27-B23)				
30	Все нормально	2,089757	МАКС(B28:B29)/B25	τ1–α = 2,6	2,62		
31							

Определяем другие статистические характеристики:
коэффициент вариации по формуле

$$V = \frac{\overline{S}}{\overline{x}} \cdot 100\%.$$

коэффициент асимметрии по формуле $g_1 = \frac{m_3}{m_2^{3/2}}.$

$$g_2 = \frac{m_4}{m_2^2} - 3.$$

Имеется также и небольшой эксцесс.
Результаты вычисления выборочных характеристик, упомянутых, выше сведены в табл.

	A	B	C	E	F	G	H	I	J
1	№ п/п	X;, см	$x - \bar{x}$	C2^2	C2^3	C2^4			
2	1	183	6,526315789	42,5928	277,974	1814,146			
3	2	170	-6,473684211	41,90859	-271,303	1756,33			
4	3	176	-0,473684211	0,224377	-0,10628	0,050345			
5	4	178	1,526315789	2,32964	3,555766	5,427222			
6	5	176	-0,473684211	0,224377	-0,10628	0,050345			
7	6	180	3,526315789	12,4349	43,84939	154,6268			
8	7	176	-0,473684211	0,224377	-0,10628	0,050345			
9	8	185	8,526315789	72,69806	619,8466	5285,008			
10	9	184	7,526315789	56,64543	426,3314	3208,705			
11	10	174	-2,473684211	6,119114	-15,1368	37,44355			
12	11	168	-8,473684211	71,80332	-608,439	5155,717			
13	12	174	-2,473684211	6,119114	-15,1368	37,44355			
14	13	189	12,52631579	156,9086	1965,487	24620,3			
15	14	172	-4,473684211	20,01385	-89,5356	400,5542			
16	15	175	-1,473684211	2,171745	-3,20047	4,716477			
17	16	167	-9,473684211	89,75069	-850,27	8055,187			
18	17	179	2,526315789	6,382271	16,12363	40,73339			
19	19	169	-7,473684211	55,85596	-417,45	3119,888			
20	20	178	1,526315789	2,32964	3,555766	5,427222			
21	X	3353		34,03878	57,1544	2826,411			
22	∑mod		88,42105263	m_2	m_3	m_4	СУММ(E2:E20)/ЧСТРОК(A2:A20)		
23	x сред	176,4737	СРЗНАЧ(B2:B21)						
24	дисперсия	35,92982	ДИСП(B2:B21)						
25	ср.кв.отклонение	5,994149	КОРЕНЬ(B24)						
26	xmin	167	МИН(B2:B21)						
27	xmax	189	МАКС(B2:B21)						
28	$\lvert x_{\min} - \bar{x} \rvert$	9,473684	ABS(B26-B23)						
29	$\lvert x_{\max} - \bar{x} \rvert$	12,52632	ABS(B27-B23)						
30	Все нормально	2,089757	МАКС(B28:B29)/B25	$\tau_{1-\alpha} = 2,6$	2,62				
31	V	3,40%	B25/B23						
32	g_1	0,287799							
33	g_2	-0,56058							

Полигон и гистограмма частот распределения

Число классов k приблизительно можно вычислить по формуле

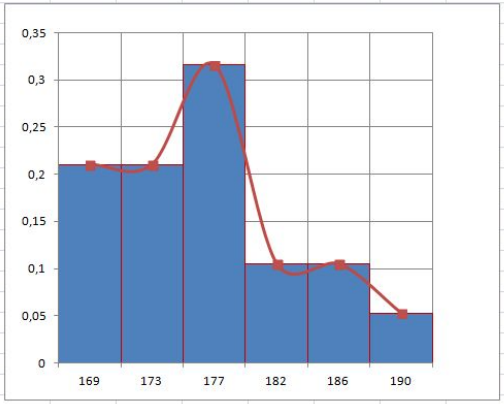
$$k \approx 1 + 3{,}32 \cdot \lg(n).$$

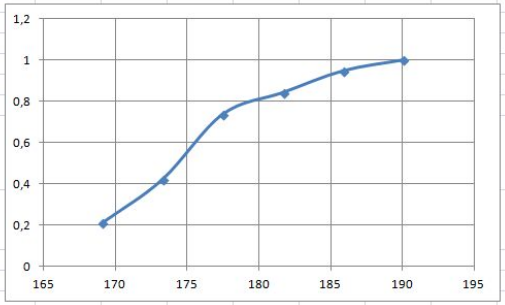
Размах варьирования по формуле $R = x_{\max} - x_{\min}$

Ширина интервалов по формуле $h = \frac{R}{k}.$

	A	B	C	D
1	№ п/п	X;, см		
2	1	183		
3	2	170		
4	3	176		
5	4	178		
6	5	176		
7	6	180		
8	7	176		
9	8	185		
10	9	184		
11	10	174		
12	11	168		
13	12	174		
14	13	189		
15	14	172		
16	15	175		
17	16	167		
18	17	179		
19	18	276		
20	19	169		
21	20	178		
22	X	3629		
23	x сред	181,45	СРЗНАЧ(B2:B21)	
24	дисперсия	529,3132	ДИСП(B2:B21)	
25	ср.кв.отклонение	23,00681	КОРЕНЬ(B24)	
26	xmin	167	МИН(B2:B21)	
27	xmax	276	МАКС(B2:B21)	
28	$\lvert x_{min} - \bar{x} \rvert$	14,45	ABS(B26-B23)	
29	$\lvert x_{max} - \bar{x} \rvert$	94,55	ABS(B27-B23)	
30	Исключить ошибку	4,109653	МАКС(B28:B29)/B25	τ1−α = 2,62;

	A	B	C	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	№ п/п	X;, см	$x - \bar{x}$	C2^2	C2^3	C2^4																				
2	1	183	6,526315789	42,5928	277,974	1814,146			167	171,1941	0	0	0	1	0	0	169									
3	2	170	-6,473684211	41,90859	-271,303	1756,33			171,1941	175,3882	1	0	0	0	0	0	173									
4	3	176	-0,473684211	0,224377	-0,10628	0,050345			175,3882	179,5823	0	0	1	0	0	0	177									
5	4	178	1,526315789	2,32964	3,555766	5,427222			179,5823	183,7764	0	0	1	0	0	0	182									
6	5	176	-0,473684211	0,224377	-0,10628	0,050345			183,7764	187,9705	0	0	1	0	0	0	186									
7	6	180	3,526315789	12,4349	43,84939	154,6268			187,9705	192,1646	0	0	0	1	0	0	190									
8	7	176	-0,473684211	0,224377	-0,10628	0,050345					0	0	1	0	0	0										
9	8	185	8,526315789	72,69806	619,8466	5285,008					0	0	0	0	1	0										
10	9	184	7,526315789	56,64543	426,3314	3208,705					0	0	0	0	1	0										
11	10	174	-2,473684211	6,119114	-15,1368	37,44355					0	1	0	0	0	0										
12	11	168	-8,473684211	71,80332	-608,439	5155,717					1	0	0	0	0	0										
13	12	174	-2,473684211	6,119114	-15,1368	37,44355					0	1	0	0	0	0										
14	13	189	12,52631579	156,9086	1965,487	24620,3					0	0	0	0	0	1										
15	14	172	-4,473684211	20,01385	-89,5356	400,5542					0	1	0	0	0	0										
16	15	175	-1,473684211	2,171745	-3,20047	4,716477					0	1	0	0	0	0										
17	16	167	-9,473684211	89,75069	-850,27	8055,187					1	0	0	0	0	0										
18	17	179	2,526315789	6,382271	16,12363	40,73339					0	0	1	0	0	0										
19	19	169	-7,473684211	55,85596	-417,45	3119,888					1	0	0	0	0	0										
20	20	178	1,526315789	2,32964	3,555766	5,427222					0	0	1	0	0	0										
21	X	3353		34,03878	57,1544	2826,411					4	4	6	2	2	1	19									
22	∑mod		88,42105263	m_2	m_3	m_4	СУММ(E2:E20)/ЧСТРОК(A2:A20)				0,210526	0,210526	0,315789	0,105263	0,105263	0,052632	1									
23	x сред	176,4737	СРЗНАЧ(B2:B21)								0,210526	0,421053	0,736842	0,842105	0,947368	1										
24	дисперсия	35,92982	ДИСП(B2:B21)	k	5	1+3,32*LOG10(ЧСТРОК(A2:A20))																				
25	ср.кв.отклонение	5,994149	КОРЕНЬ(B24)	R	22	B27-B26																				
26	xmin	167	МИН(B2:B21)	h	4,194102	F25/F24																				
27	xmax	189	МАКС(B2:B21)																							
28	$\lvert x_{min} - \bar{x} \rvert$	9,473684	ABS(B26-B23)																							
29	$\lvert x_{max} - \bar{x} \rvert$	12,52632	ABS(B27-B23)																							
30	Все нормально	2,089757	МАКС(B28:B29)/B25	τ1−α = 2,6	2,62																					
31	V	3,40%	B25/B23																							
32	g_1	0,287799																								
33	g_2	-0,56058																								
34																										
35																										





Выполняем проверку выборки на нормальность распределения по следующим критериям:

– по среднему абсолютному отклонению САО

$$CAO = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

условие соответствия по

$$\left|\frac{CAO}{\bar{S}} - 0{,}7979\right| < \frac{0{,}4}{\sqrt{n}}.$$

Условие соответствия выполняется, следовательно, гипотеза нормальности распределения выборки данных, приведенных в табл. 4, подтверждается.

Среднее значение теоретического распределения по

$$\mu = \bar{x} \pm (0{,}71 \div 0{,}6) \cdot CAO.$$

Коэффициент (0,71 ч 0,6) зависит от величины выборки n (в данном случае n = 15÷20) и 1 - α = 0,95.

Коэффициенты для определения 95%-х доверительных границ для среднего значения по САО приведены в табл. А2

$$\mu = 176{,}47 \pm 0{,}62 * 4{,}654 \text{ (см)}.$$

по размаху варьирования

$$\frac{R}{\overline{S}} = \frac{22}{5{,}994} = 3{,}67.$$

$$\frac{R}{\overline{S}}$$

$$k_{i'} \le \frac{R}{\overline{S}} \le k_{\hat{a}}.$$

При n = 19 и α = 0,10 нижняя и верхняя границы по табл. А3 соответственно равны 3,25 и 4,27, т. е.

Следовательно, гипотеза нормальности распределения подтверждается и по этому критерию

по коэффициентам асимметрии и эксцесса.

Условия:

$$G_1 \le 3S_{G1},\ G_2 \le 5S_{G2}$$

где G1 - несмещенная оценка для показателя асимметрии

$$S_{G_1} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}};$$

$$S_{G_2} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}.$$

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1;$$

$$G_2 = \frac{n-1}{(n-2)(n-3)}\left[(n+1)g_2 + 6\right].$$

Выполнение указанных условий свидетельствует, что гипотеза нормальности распределения может быть принята.

по критерию χ^2 (свойства кривой нормального распределения) должно выполняться условие

$$y = f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \cong 0{,}4e^{\frac{-z^2}{2}}.$$

где Bj — абсолютная частота в классе

$$\bar{x}_{ожид} = \frac{1}{n}\sum_{j=1}^{n_{кл}} B_j x_j;$$

4
4
6
2
2
1

$$z_j = \frac{\left|x_i - \bar{x}_{ожид}\right|}{\bar{S}_{ожид}}$$

Ej — ожидаемая частота по кривой нормального распределения .

$$E_j = f(z_j)\cdot k',$$

$$k' = \frac{nn_{\bar{s}}}{\bar{S}}, \qquad f(z_j) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_j^2}{2}}; \; \bar{S}_{ожид} \sqrt{\frac{\sum_{j=1}^{n_{кл}} B_j x_j^2 - \frac{(\sum_{j=1}^{n_{кл}} B_j x_j)^2}{n}}{n-1}};$$

В табл. 7 критерий χ^2 = 3,80. По табл. А4 находят табличное значение: 4,06

Таким образом, гипотеза о том, что наблюдаемые частоты распределены нормально, принимается на 10%-м уровне.

Вывод: так как условия соответствия на нормальность распределения выполняются, то распределение величин идет по нормальному закону. Гипотеза нормального распределения на достаточно «жестком» 10%-м уровне принимается. 95 % всех значений выборки варьируется в пределах от 173,58 до 179,36.