Проверка гипотез в MS Excel. Параметрические и непараметрические методы

СОДЕРЖАНИЕ

* Применение стандартных функций MS Excel для решения задач проверки гипотез. Доверительные интервалы для оценок среднего и дисперсии. Эмпирический тест на нормальность. Проверка гипотезы о равенстве среднего заданному значению. Проверка гипотезы о распределении по критерию согласия хи-квадрат.
* Применение пакета анализа для решения задач проверки гипотез. Проверка гипотезы о принадлежности двух дисперсий одной генеральной совокупности. Проверка гипотезы о равенстве средних при неравных дисперсиях и при равных дисперсиях. Построение таблицы рангов.

обозначения

x,y, – случайные величины, xi,yj – их выборочные значения, N – число выборок, k – число интервалов группирования, xij – двумерная матрица выборочных значений;

M - выборочное среднее, S – выборочное среднеквадратичное отклонение, S2 - выборочная дисперсия;

eM, eS – ошибки выборочных среднего и среднеквадратичного;

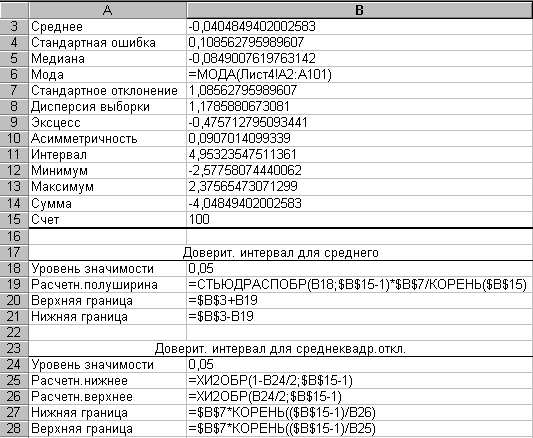
1-α - доверительная вероятность, θн – нижняя, θв – верхняя границы доверительного интервала оцениваемого параметра θ;

N{M,S} – нормальное распределение, UN{M,S} – равномерное распределение со средним M и среднеквадратичным отклонением S, T{} – распределение Стъюдента, F{} – распределение Фишера, XI{} – распределение хи-квадрат.

Практические задания по решению задач проверки гипотез

| Основные возможности в примерах и заданиях и их реализация |
| --- |
| Применение стандартных функций MS Excel для решения задач проверки гипотез |
| **Пример 1.** Проведите анализ данных в рамках задачи проверки гипотезы о значимости полученных оценок среднего и дисперсии выборки, постройте доверительные интервалы для оценок среднего и дисперсии c использованием средства **Вставка функций**. ***Применение***. Проверка гипотез о значимости полученных оценок параметров распределений основана на построении доверительных интервалов для оценок с заданным уровнем значимости α (обычно 0,05). Нулевой гипотезой считается тот факт, что истинное значение параметра попадает в построенный доверительный интервал. Доверительный интервал для среднего M: M-T{α, m}S/√N, M+ T{α, m}S/√N, где T{α, M}– табличное значение распределения Стьюдента с m=N-1 степенями свободы и уровнем значимости α. Доверительный интервал для среднеквадратичного S с заданным уровнем значимости α: S√(N-1)/XI{1-α/2, m}, S√(N-1)/XI{α/2, m}, где XI{…, m }– табличные значения распределения хи-квадрат с m=N-1 степенями свободы для доверительных вероятностей 1-α/2 и α/2.   1. Запустите MS Excel: **Пуск / Программы / Microsoft Excel** и откройте рабочую книгу **Примеры\_стат**: **Файл / Открыть /** раскройте вашу рабочую папку / двойным щелчком по имени **Примеры\_стат** загрузите рабочую книгу / перейдите на лист **Статистика N{0,1},100.** 2. При помощи средства **Вставка функций** постройте по расчетным значениям среднего и дисперсии на листах **Статистика N{0,1},100** и **Статистика N{0,1},1000** доверительные интервалы для среднего и среднеквадратичного отклонения, следуя примерному **Образцу 1**. Проанализируйте зависимость полученных доверительных пределов для среднего и среднеквадратичного отклонения от размера выборки. 3. Закройте книгу **Примеры\_стат**: **Файл / Закрыть / да** в диалоге запроса о сохранении изменений. |
| **Задание 1.** В рабочей книге **Задания\_стат** при помощи средства **Вставка функций** постройте по расчетным значениям среднего и дисперсии на листах **Статистика UN{0,1/√3},100**, **Статистика UN{0,1/√3},500** доверительные интервалы для среднего и среднеквадратичного отклонения по схеме **Примера 1**. Проанализируйте зависимость полученных доверительных пределов для среднего и среднеквадратичного отклонения от размера выборки. Закройте книгу **Задания\_стат**. |

**Образец 1**

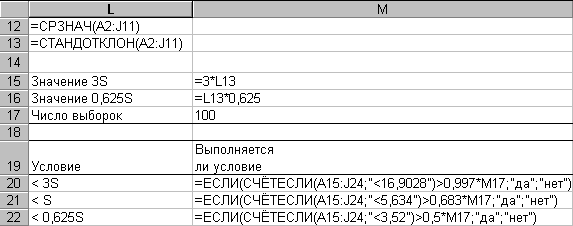


| Основные возможности в примерах и заданиях и их реализация |
| --- |
| **Пример 2.** Проведите анализ данных в рамках задачи проверки гипотезы о распределении при помощи эмпирического теста на нормальность c использованием средства **Вставка функций** MS Excel. ***Применение***. Эмпирический тест проверяет нулевую гипотезу о принадлежности распределения выборки к нормальному в соответствии со следующим алгоритмом. Рассчитывается среднее и среднеквадратичное отклонение выборочных значений и абсолютные значения отклонений выборочных значений от среднего, а затем проверяется выполнение условий:   * 99,7% отклонений от среднего меньше 3S; * 68,3% отклонений меньше S; * 50% отклонений меньше 0,625S.  1. Создайте новую рабочую книгу инструментом **Создать**  и сохраните ее под именем **Примеры\_гипотезы** в вашу рабочую папку: **Файл / Сохранить как /** раскройте папку и введите в поле **Имя файла** название книги **Примеры\_ гипотезы /** кнопка **Сохранить**. 2. Переименуйте ярлычок рабочего листа **Лист 1** в **Тест\_норм**. 3. Сформируйте массив исходных данных результатов 100 замеров отклонений от номинального размера изделия: откройте учебный файл **Образцы\_стат** командой **/ Файл / Открыть** / раскройте папку **D:\Lesson\Statistics** / двойной щелчок по файлу **Образцы\_стат** / на листе **Образец 2\_1** выделите диапазон ячеек **А1:J11** / **Правка / Копировать** / перейдите в книгу **Примеры\_гипотезы** командой **Окно / Примеры\_гипотезы** / вставьте содержимое буфера в позицию начиная с ячейки **А1** листа **Тест\_норм** командой **Правка / Вставить**. 4. Выполняя нижеследующие пункты, ориентируйтесь по **Образцу 2**. 5. Используя средство **Вставка функции**, рассчитайте среднее и среднеквадратичное отклонение, разместив расчетные формулы в ячейках **L12,L13** соответственно. 6. Рассчитайте массив отклонений выборочных значений от среднего: активизируйте ячейку **А15** / инструмент **Вставка функции** / в поле **Категории** выберите **Математические** / в поле **Функция** найдите и выберите **ABS** / **ОК** / установите курсор в поле **Число** ивведите ссылку **А2** / введите знак разности **-** / введите ссылку **L12** / перейдите к абсолютной ссылке клавишей **F4** / **ОК** / используя маркер автозаполнения ячейки **А15**, растяните формулу в ячейки **В15: J15** / не снимая выделение с диапазона **А15: J15,** растяните маркером автозаполнения выделенного диапазона формулы в ячейки **А16:J24**. |

Образец 2

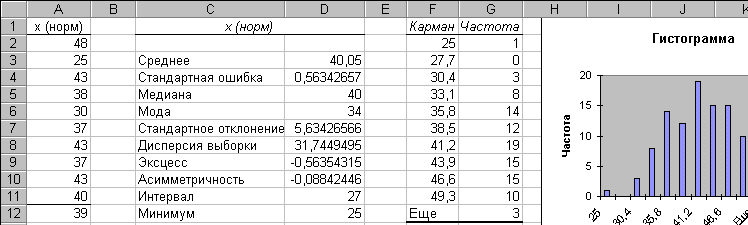
| Основные возможности в примерах и заданиях и их реализация |
| --- |
| 1. Сформируйте таблицу проверки условий эмпирического теста на нормальность и вставьте расчетные формулы согласно **Образцу 3**, например, при вставке формулы проверки первого условия <3S эмпирического теста в ячейке **M20**: активизируйте ячейку / инструмент **Вставка функции**  / в поле **Категории** выберите **Логические** / в поле **Функция** найдите и выберите **ЕСЛИ** / **ОК** / в поле **Значение\_если\_истина** напечатайте **да /** в поле **Значение\_если\_ложь** напечатайте **нет** / установите курсор в поле **Логическое выражение** / перейдите к латинской раскладке клавиатуры и напечатайте **>0,997\*M17** / установите курсор в начало этой записи перед знаком неравенства **>** / кнопка списка инструмента выбора функций / выберите **Другие функции** / выберите из категории **Статистические** функцию **СЧЕТЕСЛИ** / в поле **Диапазон** укажите ссылку на диапазон **А15:J24** / вполе **Условие** напечатайте знак **<** и расчетное значение величины 3S из ячейки **М15 / ОК** / аналогично постройте или скопируйте и отредактируйте формулы в ячейках **М21, М22.** |

Образец 3



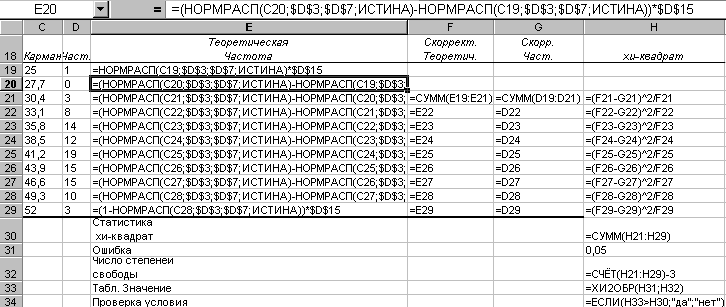
| Основные возможности в примерах и заданиях и их реализация |
| --- |
| 1. Проинтерпретируйте полученные результаты: в случае не выполнения хотя бы одного из условий эмпирического теста необходима дополнительная проверка исходной гипотезы о нормальности при помощи, например, критерия согласия хи-квадрат. При выполнении всех трех условий гипотеза о нормальном законе распределения исходных данных принимается. |
| **Задание 2.** Создайте новую рабочую книгу и сохраните ее в вашей рабочей папке под именем **Задания\_ гипотезы**.С использованием средства **Вставка функций** MS Excel проведите на листе **Задание 1** эмпирический тест на нормальность по схеме **Примера 2** следующих данных измерений размера партии 40 изделий: 19, 22, 9, 18, 18, 11, 8, 9, 17, 20, 22, 20, 21, 19, 21, 21, 20, 10, 11, 22, 23, 9, 21, 19, 12, 23, 13, 15, 13, 20, 17, 19, 20, 20, 20, 20, 19,14, 17, 17. |
| **Пример 3.** Проведите анализ данных в рамках задачи проверки гипотезы о распределении при помощи критерия согласия хи-квадрат c использованием средств **Вставка функций** и **Анализ данных** MS Excel. ***Применение***. Тест хи-квадрат проверяет нулевую гипотезу о принадлежности выборки к конкретному типу распределения (например, нормальному). При формировании статистики следует соблюдать следующие условия: число интервалов группирования больше 5, теоретическая частота попадания в интервал не менее 5. Если теоретическая частота меньше 5, данный интервал следует объединить с соседним справа. Расчетное (критериальное) значение статистики   k XI=Σm=1 (hm-Npm)2/Npm  hm – эмпирические, pm – теоретические частоты для m-го интервала сравнивается с табличным (критическим ) значением распределения хи-квадрат Пирсона с заданным уровнем значимости α (обычно 0,05) и числом степеней свободы k\*=k-r-1, где k – число интервалов группирования в таблице частот (карманов гистограммы), r – число предварительно оцениваемых параметров распределения (например, среднее и дисперсия – 2). Если критериальное значение меньше критического, нулевая гипотеза принимается, если больше или равна, гипотеза отвергается.   1. Переименуйте рабочий лист **Лист 2** книги **Примеры\_ гипотезы** в **Тест\_хи-квадрат**. В ячейке **А1** введите заголовок столбца данных **х, норм.** 2. Скопируйте исходные данные из диапазона ячеек листа **А2:J11** **Тест\_норм** на лист **Тест\_хи-квадрат** в позицию начиная с ячейки **А2**. В ячейку **А1** введите заголовок **х (норм)**. 3. Реорганизуйте скопированный массив данных на листе **Тест\_хи-квадрат** при помощи приемов перемещения диапазонов ячеек так, чтобы данные размещались в одном столбце **А**: выделите ячейки **В2:В11 / Правка / Вырезать** / активизируйте ячейку **А12** и **Правка / Вставить** / аналогично выполните для столбцов **С-J**. Результирующий массив чисел должен занимать диапазон ячеек **А2:А101**. 4. При помощи средства **Сервис / Анализ данных** рассчитайте по исходным данным на листе **Тест\_хи-квадрат** описательную статистику и постройте таблицу частот и гистограмму, следуя примерному **Образцу 4**. |

Образец 4



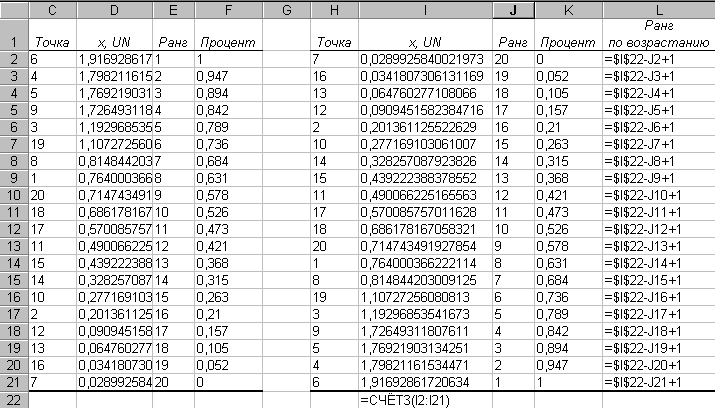
|  |
| --- |
| 1. Используя построенную таблицу частот и рассчитанные среднее и среднеквадратичное, а также стандартные встроенные функции, сформируйте таблицу для расчета статистики хи-квадрат по **Образцу 5.** Обратите внимание на ввод максимального значения вместо текста **Еще** в исходной таблице частот (ячейка **С29**) и расчетных формул теоретических частот для самого нижнего и самого верхнего интервала группирования. Эти действия необходимы для корректного вычисления теоретических частот. В столбцах скорректированных теоретических и эмпирических частот выполнено объединение тех карманов, где значение теоретических частот менее 5. Это условие правильного применения критерия. Расчетная формула для числа степеней свободы распределения хи-квадрат определяется разностью числа карманов (с учетом их объединения) и числа 3 (=количеству налагаемых связей плюс 1). |

Образец 5



|  |
| --- |
| **Задание 3.** С использованием средств **Вставка функций**  и **Анализ данных** MS Excel на листе **Задание 2**  рабочей книги **Задания\_гипотезы** проведите тест на нормальность по критерию согласия хи-квадрат по схеме **Примера 3** для данных из **Задания 1**. |
| **Пример 4**. Проведите анализ данных в рамках задачи проверки гипотезы о равенстве среднего некоторому определенному значению (для данных, взятых из нормально распределенной генеральной совокупности) c использованием средства **Вставка функций** MS Excel.  ***Применение****.* Используется для нормально распределенных независимых данных. Проверяется нулевая гипотеза о равенстве среднего определенному значению. Расчетная статистика t = (M – A)N1/.2/S, где A – предполагаемая величина среднего, сравнивается с табличным критическим значением распределения Стъюдента с уровнем значимости α/2и числом степеней свободы N-1, где N – число выборочных значений. Если расчетное значение меньше критического, нулевая гипотеза принимается, если больше, гипотеза отвергается.   1. Перейдите на рабочий лист **Лист 3** книги **Примеры\_ гипотезы** и переименуйте его в **Тест\_среднее.** 2. Сформируйте массив данных, используя функцию генерации случайных чисел средства **Анализ данных** для нормального распределения со средним M=0,15 и стандартным отклонением S=0,1 при числе выборок n=20. 3. Рассчитайте среднее и дисперсию выборки, используя стандартные функции MS Excel. 4. Сформируйте таблицу для проверки критерия, исходя из того, что критериальное значение вычисляется по формуле  t = (M – A)N1/.2 / S , где A – предполагаемая величина среднего (например, 0,09). Гипотеза о равенстве среднего константе отвергается, если по абсолютной величине критериальное значение больше верхней α/2-% точки распределения Стъюдента: |tкритериальное|> T{α/2,N-1}. ***Подсказка*.** Для расчета распределения Стъюдента используйте функцию **СТЪЮДРАСПОБР().** |
| **Задание 4.** С использованием средств **Вставка функций** MS Excel выполните на листе **Задание 3**  рабочей книги **Задания\_гипотезы** проверку гипотезы о равенстве среднего значению 1,5 по схеме **Примера 4** для 50 сгенерированных чисел, распределенных по нормальному закону со средним 1 и стандартным отклонением 3. |
| Применение пакета анализа MS Excel для решения задач проверки гипотез |
| **Пример 5**. Проведите анализ данных в рамках задачи проверки гипотезы о принадлежности двух дисперсий одной генеральной совокупности (следовательно, их равенстве) по критерию Фишера c использованием средства **Анализ данных** MS Excel. ***Применение****.* Используется для двух нормально распределенных независимых выборок данных. Проверяется нулевая гипотеза о равенстве дисперсий двух выборок размером N1, N2 - H0: S12=S22. Альтернативная гипотеза H1: S12>S22 . Расчетная статистика F= S12/S22 сравнивается с табличным критическим значением распределения Фишера Fα (односторонний критерий) с уровнем значимости αи числом степеней свободы N1-1, N2-1. Если расчетное значение меньше критического F<Fα, нулевая гипотеза принимается. Если статистика F меньше единицы, односторонний критерий не работает, и нулевая гипотеза принимается при выполнении условия F1-α/2<F<Fα/2  двухстороннего критерия с уровнем значимости α/2.   1. Перейдите на рабочий лист **Лист 4** книги **Примеры\_ гипотезы** и переименуйте его в **Тест\_рав\_дисп.** 2. Скопируйте массив исходных данных двух независимых выборок значений погрешностей измерения прибора из учебного файла **Образцы\_стат** с листа **Образец 2\_2**. 3. Запустите процедуру проверки гипотезы: **Сервис / Анализ данных / Двухвыборочный тест для дисперсии /**  в одноименном окне укажите диапазоны ячеек для 1 и 2 выборок в полях **Интервал переменной …**, введите уровень значимости **0,05** в поле **Альфа**, укажите верхнюю левую ячейку размещения результатов в поле **Выходной интервал** и **ОК**. 4. Проанализируйте полученные результаты. По условиям критерия нулевая гипотеза отвергается, если значение F статистики Фишера больше верхнего критического или меньше нижнего. 5. Постройте для обеих выборок гистограммы с полиномиальным трендом. |
| **Задание 5.** С использованием средства **Анализ данных** MS Excel по схеме **Примера 5** проведите на листе **Задание 4** рабочей книги **Задания\_гипотезы** проверку гипотезы о принадлежности двух дисперсий одной генеральной совокупности по критерию Фишера для двух выборок данных измерения опухоли: 0,027 0,036 0,1 0,12 0,32 0,46 0,049 0,105 0,075 0,4 0,08 0,105 0,075 0,12 0,06 0,1050,0128 |
| **Пример 6**. Проведите анализ данных в рамках задачи проверки гипотезы о равенстве средних при неравных дисперсиях и объемах выборок по критерию Стъюдента c использованием средства **Анализ данных** MS Excel. ***Применение****.* Используется для двух нормально распределенных независимых выборок данных размером N1, N2 с различными дисперсиями. Проверяется нулевая гипотеза о равенстве средних двух выборок H0: М1=М2. Альтернативная гипотеза H1: М1≠М2. Расчетная статистика t=( М1-М2)/ (S12/N1+S22/N2)1/2 сравнивается с табличным критическим значением распределения Стъюдента с уровнем значимости α/2и числом степеней свободы V, где V=(S12/N1+S22/N2)2/[(S12/N1)2/(N1+1)+(S22/N2)2/(N2+1)]-2. Если расчетное значение по абсолютной величине меньше критического, нулевая гипотеза принимается.   1. Перейдите на рабочий лист **Лист 5** книги **Примеры\_ гипотезы** и переименуйте его в **Тест\_рав\_сред1.** 2. Сформируйте массив данных, используя функцию генерации случайных чисел средства **Анализ данных** для нормального распределения со средним 1 и стандартным отклонением 1 и 2 для двух выборок объемом 15 и 20 чисел соответственно, разместив их в столбцах **В** и **С**. 3. Запустите процедуру проверки гипотезы: **Сервис / Анализ данных / Двухвыборочный t-тест с различными дисперсиями /** в одноименном окне укажите диапазоны ячеек для 1 и 2 выборок в полях **Интервал переменной …**, введите уровень значимости **0,05** в поле **Альфа**, укажите верхнюю левую ячейку размещения результатов в поле **Выходной интервал** и **ОК**. 4. Проанализируйте полученные результаты. По условиям критерия нулевая гипотеза отвергается, если значение t-статистики Стъюдента по абсолютной величине больше верхней точки распределения или критического значения. 5. Постройте для обеих выборок гистограммы с трендом скользящего среднего. |
| **Задание 6.** С использованием средства **Анализ данных** MS Excel по схеме **Примера 6** проведите на листе **Задание 5** рабочей книги **Задания\_гипотезы** проверку гипотезы о равенстве средних при неравных дисперсиях по сгенерированным данным для нормального распределения со средним 5 и стандартным отклонением 2 и 4 для двух выборок объемом 20 и 25 чисел. |
| **Пример 7**. Проведите анализ данных в рамках задачи проверки гипотезы о равенстве средних при равных дисперсиях по критерию Стъюдента c использованием средства **Анализ данных** MS Excel. ***Применение****.* Используется для двух нормально распределенных независимых выборок данных размером N1 и N2 с одинаковыми дисперсиями. Проверяется нулевая гипотеза о равенстве средних двух выборок против альтернативной гипотезы о их неравенстве (двухсторонний критерий). Расчетная статистика  t=(М1-М2)/{[(S12(N1-1)+S22(N2-1)](1/N1+1/N2)/(N1+N2 –2)}1/2 сравнивается с табличным критическим значением распределения Стъюдента с уровнем значимости α/2и числом степеней свободы N1+N2 –2. Если расчетное значение по абсолютной величине меньше критического, нулевая гипотеза принимается.   1. Перейдите на рабочий лист **Лист 6** книги **Примеры\_ гипотезы** и переименуйте его в **Тест\_рав\_сред.2** 2. Скопируйте массив исходных данных двух независимых серий испытаний лазера из учебного файла **Образцы\_стат** с листа **Образец 2\_3**. 3. Запустите процедуру проверки гипотезы: **Сервис / Анализ данных / Двухвыборочный t-тест с одинаковыми дисперсиями /** в одноименном окне укажите диапазоны ячеек для 1 и 2 выборок в полях **Интервал переменной …**, введите уровень значимости **0,05** в поле **Альфа**, укажите верхнюю левую ячейку размещения результатов в поле **Выходной интервал** и **ОК**. При вводе значения в поле **Гипотетическая средняя разность** проверяется гипотеза о разности значений средних двух выборок. 4. Проанализируйте полученные результаты и примите решение. 5. Постройте для обеих выборок гистограммы с полиномиальным трендом. |
| **Задание 7.** С использованием средства **Анализ данных** MS Excel по схеме **Примера 6** проведите на листе **Задание 6** рабочей книги **Задания\_гипотезы** проверку гипотезы о равенстве средних при равных дисперсиях по сгенерированным данным для нормального распределения со средним 2 и стандартным отклонением 2 для двух выборок объемом 40 точек каждая. |

Образец 6



|  |
| --- |
| **Пример 8**. Проведите построение таблицы рангов и персентилей c использованием средства **Анализ данных** MS Excel. ***Рангом*** выборочного значения называется номер, который получит это значение в упорядоченной совокупности всех выборочных значений после их упорядочения по некоторому правилу (по возрастанию или по убыванию).   1. Перейдите на рабочий лист **Лист 7** книги **Примеры\_ гипотезы** и переименуйте его в **Ранги.** Введите в ячейке **А1** заголовок столбца **х,UN**. 2. При помощи средства **Анализ данных** сгенерируйте массив исходных данных с равномерным распределением 1-ой переменной на 20 точек от 0 до 2 в диапазоне ячеек **А2:А21**. Уменьшите разрядность представления чисел до четырех знаков после запятой. 3. Запустите процедуру построения таблицы рангов: **Сервис / Анализ данных / Ранг и персентиль /** в одноименном окне укажите диапазон ячеек **А1:А21** в поле **Входной** **Интервал**, включите переключатель **По столбцам**, установите флажок **Метки в первой строке**, укажите верхнюю левую ячейку размещения результатов **С1** в поле **Выходной интервал** и **ОК**. 4. Проанализируйте полученные результаты: в столбце **Точка** отображается порядковый номер в исходной последовательности данных данного в столбце **х,UN** значения; в столбце **Ранг** приводится порядковый номер соответствующего значения в упорядоченной **по убыванию** последовательности (вариационном ряде), в столбце **Процент** рассчитывается относительное положение данного значения в массиве данных в процентах. ***Замечание***. Процедура корректно работает только для массивов данных, не содержащих повторяющиеся значения. 5. Ориентируясь по **Образцу 6**, скорректируйте полученные значения рангов для упорядоченного **по возрастанию** вариационного ряда: скопируйте полученную таблицу и вставьте ее дубль, начиная с ячейки **H1 /** отсортируйте дубль таблицы в порядке возрастания значений в столбце **х,UN** / выделите таблицу целиком, включая заголовки столбцов / **Данные / Сортировка /** кнопка **Параметры** / а диалоге **Параметры сортировки** в разделе **Сортировать** включите переключатель **строки диапазона / ОК /** в разделе **Идентифицировать поля по** включите переключатель **подписям** / в поле **Сортировать по** выберите заголовок столбца **х,UN** и включите переключатель **по возрастанию / ОК /** в ячейку **I22** под последним значением строкой таблицы в столбце **х,UN** введите вспомогательную формулу для подсчета количества значений вариационного ряда примерного вида **=СЧЁТЗ(I2:I21)** / в ячейке **L2** строки заголовков и столбца справа от столбца **Процент** введите заголовок столбца для расчета скорректированного ранга **Ранг по возрастанию** / переместитесь на ячейку ниже в этом столбце и введите формулу для пересчета значения рангавида **=$I$22-J2+1** / растяните формулу в нижние ячейки данного столбца. 6. Постройте для выборки гистограмму с линейным трендом. |
| **Задание 8.** С использованием средства **Анализ данных** MS Excel по схеме **Примера 7** постройте на листе **Задание 7** рабочей книги **Задания\_гипотезы** таблицу рангов по сгенерированным данным для нормального распределения со средним 2 и стандартным отклонением 3 для 1 переменой объемом 20 точек. Постройте для выборки гистограмму с полиномиальным трендом. |