Лекция 10

Электронный процессор Excel. (продолжение)

§5.11. Анализ данных

Набор средств анализа данных (так называемый пакет анализа), предназначенный для решения сложных статистических и инженерных задач. Для анализа данных с помощью этих инструментов следует указать входные данные и выбрать параметры; анализ будет выполнен с помощью подходящей статистической или инженерной макрофункции, а результат будет помещен в выходной диапазон. Другие средства позволяют представить результаты анализа в графическом виде.

Также в Microsoft Excel представлено большое число статистических, финансовых и инженерных функций. Некоторые из них являются встроенными, другие доступны только после установки пакета анализа.

Сервис / анализ данных

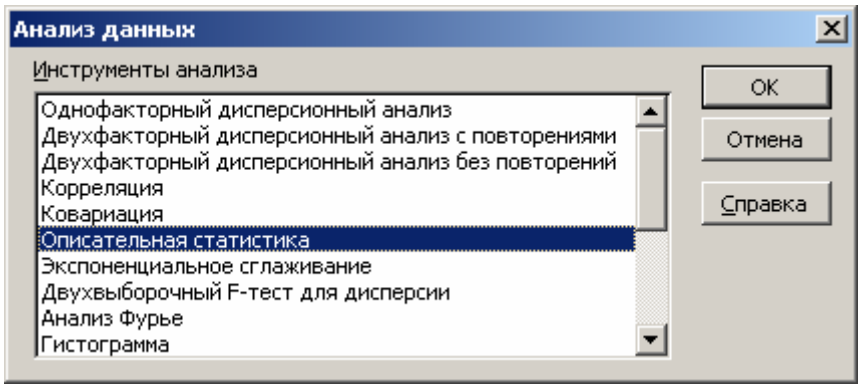

рис.5.25.

<u>Дисперсионный анализ</u>

Существует несколько видов дисперсионного анализа. Требуемый вариант выбирается с учетом числа факторов и имеющихся выборок из генеральной совокупности.

Однофакторный дисперсионный анализ.

Двухфакторный дисперсионный анализ с повторениями.

Двухфакторный дисперсионный анализ без повторения.

<u>Ковариационный анализ</u>

<u>Корреляционный анализ</u>

Описательная статистика.

Это средство анализа служит для создания одномерного статистического отчета, содержащего информацию о центральной тенденции и изменчивости входных данных.

Рассмотрим методы предварительной статистической обработки данных. Предполагаем, что все статистические данные однородны; для дальнейшей обработки существуют следующие возможности:

1) с использованием функций листа;

2) с использованием программ на встроенном языке программированиея VBA;

3) различных надстроек, например, «Пакет анализа».

При использовании описательной статистики автоматически генерируется отчет со следующими параметрами:

1) средняя величина или математическое ожидание – произведение суммы всех возможных значений случайной величины на вероятность этих значений или абсцисса центра тяжести в случае, когда масса распределена по оси абсцисс непрерывно с плотностью f(x);

$\bar{y} = M(y) = \Sigma y_i / N$,

$\bar{y}$ – средняя величина; M(y) – математическое ожидание величины y; N – количество значений.

Модой случайной величины называется её наиболее вероятное значение. Если кривая распределения имеет более одного максимума, то распределение называется «полимодальным». Распределение обладающее не максимумом, а минимумом называется «антимодальным».

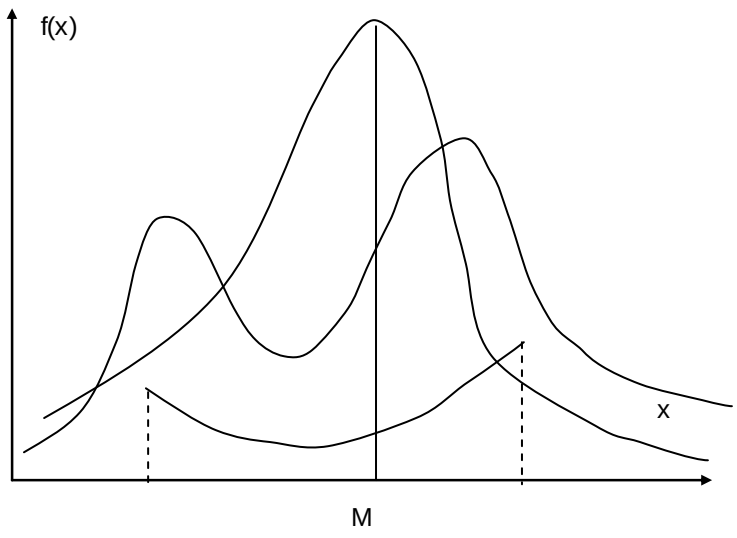

рис. 5.26.

Медианой случайной величины X называется её значение M_e, для которой одинаково вероятно, окажется ли случайная величина меньше или больше M_e (геометрически – это абсцисса точки в которой площадь, ограниченная кривой распределения, делится пополам).

$$\bar{y} = \frac{y_{\max} - y_{\min}}{2}$$

Дисперсией называют сумму отклонений значений от наиболее вероятного значения. Различают два вида дисперсии:

$$D = D(y) = \frac{\sum (y_i - \bar{y})^2}{N}$$

$$D_0 = D_0(y) = \frac{\sum (y_i - \bar{y})^2}{N-1}$$

D называется смещенная дисперсия или смещенная оценка;

D_0 – несмещенная дисперсия (для величин, распределенных по нормальному закону).

Под стандартным отклонением понимается среднеквадратичное отклонение:

$$\sigma = \sqrt{D}\,,\ \sigma_0 = \sqrt{D_0}\,.$$

Стандартное отклонение показывает среднеквадратичную погрешность при оценке значений переменной по среднему.

Эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение. Отрицательный эксцесс обозначает относительно сглаженное распределение.

$$E = \frac{1}{ND^2}\sum(y_i - \bar{y})^4$$ (для нормального распределения стремится к 3).

Асимметрия характеризует степень несимметричности распределения относительно его среднего. Положительная асимметрия указывает на отклонение распределения в сторону положительных значений. Отрицательная асимметрия указывает на отклонение распределения в сторону отрицательных значений.

$$\mathrm{A} = \frac{1}{ND^{3/2}}\sum(y_i - \bar{y})^3$$ (при нормальном распределении стремиться к 0).

Примечание: особенность нормального закона распределения (распределения Гаусса) является равенство среднего значения математическому ожиданию величины. Для остальных распределений среднее значение не является наиболее вероятным. Эксцесс и асимметрия характеризуют степень близости наблюдаемого распределения к нормальному.

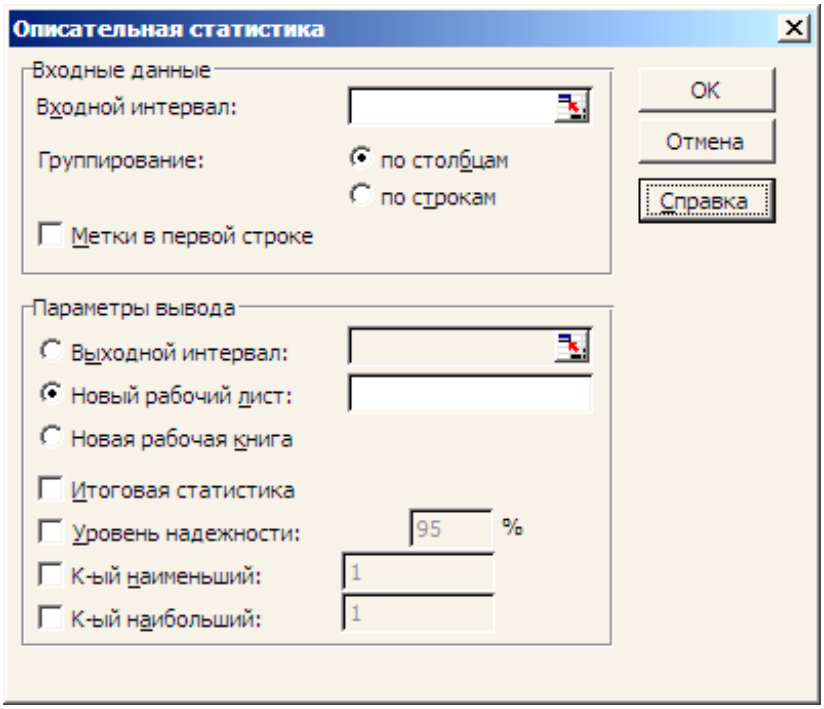

рис.5.27.

Элементы диалогового окна «Описательная статистика»

Некоторые сведения в этом разделе могут быть неприменимы к отдельным языкам.

Входной диапазон: ссылка на диапазон, содержащий анализируемые данные. Ссылка должна состоять не менее чем из двух смежных диапазонов данных, данные в которых расположены по строкам или столбцам.

Группирование: установите переключатель в положение По столбцам или По строкам в зависимости от расположения данных во входном диапазоне.

Метки в первой строке/Метки в первом столбце: если первая строка исходного диапазона содержит названия столбцов, установите переключатель в положение Метки в первой строке. Если названия строк находятся в первом столбце входного диапазона, установите переключатель в положение Метки в первом столбце. Если входной диапазон не содержит меток, то необходимые заголовки в выходном диапазоне будут созданы автоматически.

Уровень надежности: установите флажок, если в выходную таблицу необходимо включить строку для уровня надежности. В поле введите требуемое значение. Например, значение 95% вычисляет уровень надежности среднего со значимостью 0.05.

К-ый наибольший: установите флажок, если в выходную таблицу необходимо включить строку для k-го наибольшего значения для каждого диапазона данных. В соответствующем окне введите число k. Если k равно 1, эта строка будет содержать максимум из набора данных.

К-ый наименьший: установите флажок, если в выходную таблицу необходимо включить строку для k-го наименьшего значения для каждого диапазона данных. В соответствующем окне введите число k. Если k равно 1, эта строка будет содержать минимум из набора данных.

Выходной диапазон: введите ссылку на левую верхнюю ячейку выходного диапазона. Этот инструмент анализа выводит два столбца сведений для каждого набора данных. Левый столбец содержит метки статистических данных; правый столбец содержит статистические данные. Состоящий их двух

столбцов диапазон статистических данных будет выведен для каждого столбца или для каждой строки входного диапазона в зависимости от положения переключателя Группирование.

Новый лист: установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенном напротив соответствующего положения переключателя.

Новая книга: установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге.

Итоговая статистика: установите флажок, если в выходном диапазоне необходимо получить по одному полю для каждого из следующих видов статистических данных: Среднее, Стандартная ошибка (среднего), Медиана, Мода, Стандартное отклонение, Дисперсия выборки, Эксцесс, Асимметричность, Интервал, Минимум, Максимум, Сумма, Счет, Наибольшее (#), Наименьшее (#), Уровень надежности.

Экспоненциальное сглаживание

Применяется для предсказания значения на основе прогноза для предыдущего периода, скорректированного с учетом погрешностей в этом прогнозе. При анализе используется константа сглаживания a, по величине которой определяется степень влияния на прогнозы погрешностей в предыдущем прогнозе.

Примечание. Для константы сглаживания наиболее подходящими являются значения от 0,2 до 0,3. Эти значения показывают, что ошибка текущего прогноза установлена на уровне от 20 до 30 процентов ошибки предыдущего прогноза. Более высокие значения константы ускоряют отклик, но могут привести к непредсказуемым выбросам. Низкие значения константы могут привести к большим промежуткам между предсказанными значениями.

Двухвыборочный F-тест для дисперсии

Двухвыборочный F-тест применяется для сравнения дисперсий двух генеральных совокупностей.

Например, можно использовать F-тест по выборкам результатов заплыва для каждой из двух команд. Это средство предоставляет результаты сравнения нулевой гипотезы о том, что эти две выборки взяты из распределения с равными дисперсиями, с гипотезой, предполагающей, что дисперсии различны в базовом распределении.

С помощью этого средства вычисляется значение f F-статистики (или F-коэффициент). Значение f, близкое к 1, показывает, что дисперсии генеральной совокупности равны. В таблице результатов, если f < 1, "P(F <= f) одностороннее" дает возможность наблюдения значения F-статистики меньшего f при равных дисперсиях генеральной совокупности и F критическом одностороннем выдает критическое значение меньше 1 для выбранного уровня значимости Alpha. Если f > 1, "P(F <= f) одностороннее" дает возможность наблюдения значения F-статистики большего f при равных дисперсиях генеральной совокупности и F критическом одностороннем выдает критическое значение большее 1 для Alpha.

Анализ Фурье

Предназначается для решения задач в линейных системах и анализа периодических данных на основе метода быстрого преобразования Фурье (БПФ). Эта процедура поддерживает также обратные преобразования, при этом, инвертирование преобразованных данных возвращает исходные данные.

Гистограмма

Используется для вычисления выборочных и интегральных частот попадания данных в указанные интервалы значений. При этом рассчитываются числа попаданий для заданного диапазона ячеек.

Например, необходимо выявить тип распределения успеваемости в группе из 20 студентов. Таблица гистограммы состоит из границ шкалы оценок и количеств студентов, уровень успеваемости которых находится между самой нижней границей и текущей границей. Наиболее часто повторяемый уровень является модой интервала данных.

Скользящее среднее

Скользящее среднее используется для расчета значений в прогнозируемом периоде на основе среднего значения переменной для указанного числа предшествующих периодов. Скользящее среднее, в отличие от простого среднего для всей выборки, содержит сведения о тенденциях изменения данных. Этот метод может использоваться для прогноза сбыта, запасов и других процессов. Расчет прогнозируемых значений выполняется по следующей формуле.

$$F_{(t+1)} = \frac{1}{N} \sum_{j=1}^{N} A_{t-j+1}$$

где: N — число предшествующих периодов, входящих в скользящее среднее; Aj — фактическое значение в момент времени j; Fj — прогнозируемое значение в момент времени j.

Генерация случайных чисел

Используется для заполнения диапазона случайными числами, извлеченными из одного или нескольких распределений. С помощью данной процедуры можно моделировать объекты, имеющие случайную природу, по известному распределению вероятностей.

Например, можно использовать нормальное распределение для моделирования совокупности данных по росту индивидуумов, или использовать распределение Бернулли для двух вероятных исходов, чтобы описать совокупность результатов бросания монеты.

Ранг и персентиль

Используется для вывода таблицы, содержащей порядковый и процентный ранги для каждого значения в наборе данных. Данная процедура может быть применена для анализа относительного взаиморасположения данных в наборе. Она использует функции РАНГ и ПРОЦЕНТРАНГ. РАНГ не работает со связанными значениями. Если требуется учитывать связанные значения, можно воспользоваться функцией РАНГ вместе с коэффициентом изменения, описанным в файле справки для функции РАНГ.

Регрессия

Линейный регрессионный анализ заключается в подборе графика для набора наблюдений с помощью метода наименьших квадратов. Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных.

Например, на спортивные качества атлета влияют несколько факторов, включая возраст, рост и вес. Регрессия пропорционально распределяет меру качества по этим трем факторам на основе его спортивных результатов. Результаты регрессии впоследствии могут быть использованы для предсказания качеств нового, непроверенного атлета.

Регрессия использует функцию ЛИНЕЙН.

Выборка

Создает выборку из генеральной совокупности, рассматривая входной диапазон как генеральную совокупность. Если совокупность слишком велика для обработки или построения диаграммы, можно использовать представительную выборку. Кроме того, если предполагается периодичность входных данных, то можно создать выборку, содержащую значения только из отдельной части цикла.

Например, если входной диапазон содержит данные для квартальных продаж, создание выборки с периодом 4 разместит в выходном диапазоне значения продаж из одного и того же квартала.

Т-тест

Двухвыборочный t-тест проверяет равенство средних значений генеральной совокупности по каждой выборке. Эти три средства допускают следующие условия: равные дисперсии генерального распределения, дисперсии генеральной совокупности не равны, а также представление двух выборок до и после наблюдения по одному и тому же субъекту.

Для всех трех средств, перечисленных ниже, значение t-статистики t вычисляется и отображается как "t-статистика" в выводимой таблице. В зависимости от данных, это значение t может быть отрицательным или неотрицательным. Если предположить, что средние генеральной совокупности равны, при t < 0 “P(T <= t) одностороннее” дает вероятность того, что наблюдаемое значение t-статистики будет более отрицательным, чем t. При t >=0 “P(T <= t) одностороннее” делает возможным наблюдение значения t-статистики, которое будет более положительным чем t. “t критическое одностороннее” выдает пороговое значение, так что вероятность наблюдения значения t-статистики большего или равного “t критическое одностороннее” равно Alpha.

“P(T <= t) двустороннее” дает вероятность наблюдения значения t-статистики по абсолютному значению большего чем t. “P критическое двустороннее” выдает пороговое значение, так что значение вероятности наблюдения значения t- статистики по абсолютному значению большего “P критическое двустороннее” равно Alpha.

Двухвыборочный t-тест с одинаковыми дисперсиями. Двухвыборочный t-тест Стьюдента служит для проверки гипотезы о равенстве средних для двух выборок. Эта форма t-теста предполагает совпадение значений дисперсии генеральных совокупностей и обычно называется гомоскедастическим t-тестом.

Двухвыборочный t-тест с разными дисперсиями. Двухвыборочный t-тест Стьюдента используется для проверки гипотезы о равенстве средних для двух выборок данных из разных генеральных совокупностей. Эта форма t-теста предполагает несовпадение дисперсий генеральных совокупностей и обычно называется гетероскедастическим t-тестом. Если тестируется одна и та же генеральная совокупность, используйте парный тест.

Для определения тестовой величины t используется следующая формула.

$$t' = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

Следующая формула используется для вычисления степени свободы df. Так как результат вычисления обычно не бывает целым числом, значение df округляется до целого для получения порогового значения из t-таблицы. Функция Excel TTECT по возможности использует вычисленные значения без округления для вычисления значения TTECT с нецелым значением df. Из-за разницы подходов к определению степеней свободы, результаты функций TTECT и t-тест будут различаться в случае с разными дисперсиями.

$$df = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{\left(S_1^2 / m\right)^2}{m-1} + \frac{\left(S_2^2 / n\right)^2}{n-1}}$$

Парный двухвыборочный t-тест для средних. Парный двухвыборочный t-тест Стьюдента используется для проверки гипотезы о различии средних для двух выборок данных. В нем не предполагается равенство дисперсий генеральных совокупностей, из которых выбраны данные. Парный тест используется, когда имеется естественная парность наблюдений в выборках,

например, когда генеральная совокупность тестируется дважды — до и после эксперимента.

Примечание. Одним из результатов теста является совокупная дисперсия (совокупная мера распределения данных вокруг среднего значения), вычисляемая по следующей формуле.

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$$

<u>Z-тест</u>

Двухвыборочный z-тест для средних с известными дисперсиями. Используется для проверки гипотезы о различии между средними двух генеральных совокупностей. При неизвестных значения дисперсий следует использовать функцию ZTECT.

При использовании функции z-тест следует внимательно просматривать результат. “P(Z <= z) одностороннее” на самом деле есть P(Z >= ABS(z)), вероятность z-значения, удаленного от 0 в том же направлении, что и наблюдаемое z-значение при одинаковых средних значениях генеральной совокупности. “P(Z <= z) двустороннее” на самом деле есть P(Z >= ABS(z) или Z <= -ABS(z)), вероятность z-значения, удаленного от 0 в том же направлении, что и наблюдаемое z-значение при одинаковых средних значениях генеральной совокупности. Двусторонний результат является односторонним результатом, умноженным на 2. Функцию z-тест можно применять для гипотезы об особом ненулевом значении разницы между двумя средними генеральных совокупностей.

Например, этот тест может использоваться для определения различия между характеристиками двух моделей автомобилей.