УДК 677.11.021

ПРИМЕНЕНИЕ СОВРЕМЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ СРЕДСТВ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ПОКАЗАТЕЛЕЙ ЛУБОВОЛОКНИСТОГО МАТЕРИАЛА

В.С.Толмачев, *соискатель*

ХЕРСОНСКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Данная статья посвящена статистическому анализу результатов полевых и технологических опытов по первичной обработке лубяных культур с помощью современных информационных технологий и программного обеспечения.

Эффективность научных исследований во многом зависит от качества обработки экспериментальных данных. Следует отметить, что в последнее время объем вычислений, сопровождающих обработку, существенно возрос. Это произошло вследствие необратимого усложнения технологий, увеличения количества рассматриваемых факторов, а также повышения требований к точности результатов.

Как известно, особенностью проведения опытов по первичной обработке лубяных культур является естественная неоднородность используемого сырья: соломы, тресты, волокна, а также влияние различных воздействий рабочих органов используемых агрегатов и технологических операций.

В связи с этим при обработке результатов таких экспериментов большое распространение получили методы математической статистики такие как:

- количественного и описательного анализа;
- первичного анализа статистических данных;
- вариационного ряда;
- выборочный;
- анализа взаимосвязей;
- многомерного статистического анализа;
- анализа качественных признаков.

Только методами математической статистики можно обосновать необходимое количество испытаний, которое нужно провести для обеспечения заданной точности? а также проконтролировать полученную продукцию (*волокно*) и проанализировать ее изменение в процессе обработки тресты, а также связать физико-механические свойства волокна (*гибкость, линейная плотность, прочность и т.д.*) с факторами технологического процесса (*частота вращения рабочих органов, скорость движения материала, амплитуда колебаний и др.*)

Для успешного осуществления вычислительной работы, повышения ее производительности и качества исследователи широко применяют электронно-вычислительные машины (ЭВМ).

Очевидно, что эффективное использование возможностей ЭВМ становится реальным только при наличии достаточно универсальных и отлаженных программ, к которым можно отнести такие программные пакеты:

- STADIA (http://www.protein.bio.msu.su/~akula/index.htm);
- STATA (http://www.stata.com/);
- STATISTICA (http://www.statsoft.com/);
- SPSS (Statistical Package for Social Science) http://www.spss.com/ ;
- JMR (http://www.jmp.com/);
- SYSTAT (http://systat.com/);
- NCSS (http://www.ncss.com/);
- MINITAB 14. (http://www.minitab.com/);
- STATGRAPHICS PLUS. (http://www.statgraphics.com/);
- PRISM (http://www.graphpad.com/);
- Microsoft Excel;
- OpenOffice.

Несмотря на такое множество программ, одной из распространенных и доступных можно считать табличный процессор Microsoft Excel из офисного программного пакета MicrosoftOffice и его аналог - программа Calk из пакета OpenOffice.

В процессе анализа экспериментальных данных с использованием информационных технологий, как правило, присутствуют несколько основных этапов:

- ввод данных;
- преобразование данных;
- визуализация данных;
- статистический анализ;
- представление результатов.

Данная статья направлена на обобщение опыта использования табличных процессоров в научно-исследовательской работе с лубяными культурами для вычисления основных статистических характеристик.

Microsoft Excel – очень мощный, достаточно универсальный табличный процессор, ориентированный на различные сферы деятельности [1].

В состав Excel входит библиотека, содержащая 78 статистических функций, ориентированных на решение самых различных задач прикладного статистического анализа. Функции библиотеки сгруппированы по категориям: финансовые, дата и время, математические, статистические, текстовые, логические, ссылки и массивы, работа с базой данных, проверка свойств и значений.

Для работы с полученными данными их нужно внести в ячейки таблицы в определенную форму (Рис.1).

Повторность	масса кипы 1	масса кипы 2	масса кипы 3
1	24,12	21,62	23,12
2	22,79	22,82	23,2
3	24,1	23,13	22,47
4	23,22	22,87	23,17
5	22,36	23,25	23,72
6	24,31	23,08	22,92
7	24,22	21,42	22,66
8	22,27	21,74	23,61
Среднее арифметическое	23,42	22,49	23,11
Максимальное значение	24,31	23,25	23,72
Минимальное значение	22,27	21,42	22,47
Среднеквадратическое отклонение	0,87	0,76	0,43

Рис.1 *– Пример таблицы с данными*

Затем необходимо поставить курсор на одну из пустых ячеек и, выбрав соответствующий пункт меню «Вставка», вставить нужную функцию (Рис.2).

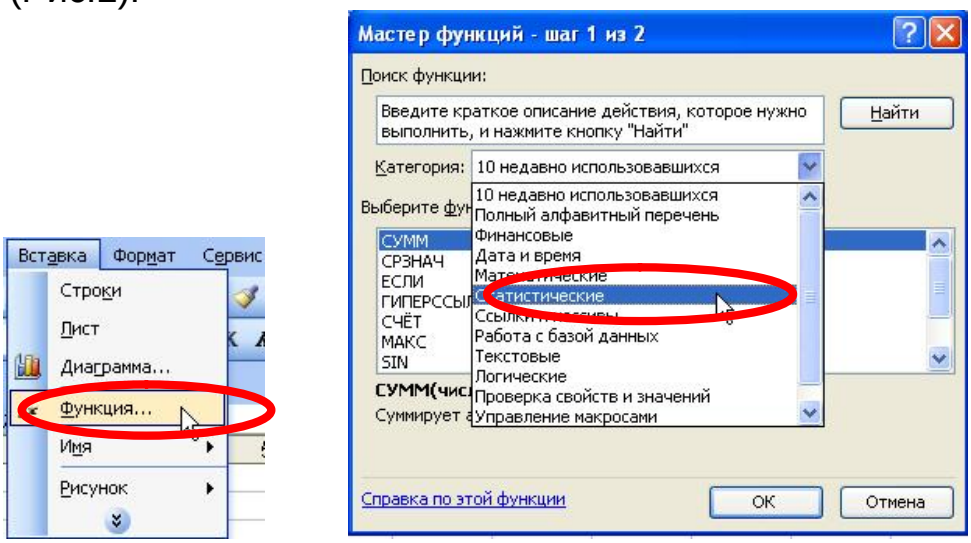

Рис.2 *– Окно запуска мастера функций*

После этого необходимо определить диапазон данных для вычисления, которые должны быть предварительно внесены в ячейки.

В Microsoft Excel мы можем найти ряд распространенных статистических функций для расчета:

- **МОДА** - определение моды;
- **МЕДИАНА** - определение медианы;
- **МИН** - определение минимального значения;
- **МАКС** - определение максимального значения;
- **СРЗНАЧ** - вычисление среднего арифметического;
- **СРОТКЛ** - определение среднего отклонения;
- **СТАНДОТКЛОН** - определение среднего квадратического отклонения;
- **СТОШҮХ** - определение стандартной ошибки;
- **КОРРЕЛ -** определение коэффициента корреляции.

Использование этих и других функций должно быть осознанным, поскольку исследователь должен четко понимать какие действия производятся над числами, и с какой целью он выбирает ту или иную функцию.

При изучении какого-либо свойства или признака на большом количестве однородных образцов исследования, например при измерении массы кип льна, урожайности соломы, семян, выхода волокна, его линейной плотности, длины, цветовых характеристик (*координат цвета*), гибкости и прочности, всегда приходится сталкиваться с определенной изменчивостью этих признаков у отдельных образцов.

Все эти и подобные им числовые результаты исследования обычно являются крайне разнообразными, варьирующими в определенных пределах даже внутри вполне однородной группы образцов исследования.

При желании сопоставить друг с другом подобные результаты измерений для нескольких разнородных групп образцов, например, при сравнении массы нескольких отобранных кип льна полученных из разных источников, массы полученного модифицированного волокна и т.д. (табл.1) [2,3], было бы совершенно бесполезно пытаться сравнивать их между собой, так как в силу неизбежных колебаний параметров внутри отобранных групп мы рисковали бы случайно подвергнуть сравнению сильно отличающиеся образцы и получить, таким образом, ложное представление о действительном соотношении.

Единственный правильный путь такого сопоставления заключается в предварительном определении некоторой общей числовой характеристики, например, типичной массы всей группы однородных образцов исследования и в сравнении друг с другом уже таких общих суммарных результатов.

Пусть, например, требуется сравнить массу модифицированного волокна, полученного из трех групп отобранного материала (табл.1). Для того, чтобы это стало вполне очевидным, необходимо вычислить типичную массу модифицированного волокна по трем группам отдельно и эти общие характеристики сопоставить уже друг с другом.

Таблица 1 *– Результаты экспериментальных исследований модифицированного лляного волокна*

№ повторности	Масса кипы, кг			Модифицированное волокно			Процент выхода волокна, %		
	1	2	3	1	2	3	1	2	3
1	58,93	58,68	58,78	24,12	21,62	23,12	40,93	36,85	39,33
2	61,25	60,75	57,94	22,79	22,82	23,2	37,21	37,56	40,05
3	60,35	61,02	59,31	24,1	23,13	22,47	39,94	37,91	37,89
4	59,8	59,36	60,24	23,22	22,87	23,17	38,83	38,54	38,46
5	60,73	62,43	61,35	22,36	23,25	23,72	36,82	37,24	38,67
6	61,52	60,38	58,64	24,31	23,08	22,92	39,52	38,23	39,08
7	60,49	58,93	60,86	24,22	21,42	22,66	40,02	36,34	37,23
8	60,9	60,56	62,32	22,27	21,74	23,61	36,57	35,89	37,89

Для этого существуют три основных приема. Первый из них заключается в выяснении величины данного признака у того образца, который занимает центральное положение среди всех остальных образцов исследования в группе.

Если расположить значения массы в ряд в порядке возрастания или убывания, то значение стоящее как раз в середине этого ряда, и может быть принято за типичную массу всех образцов этой группы.

Величина признака, полученная этим способом, носит название «медианы». Если середина такого ряда придется в промежутке между двумя массами (*при четном их количестве*), то за медиану принимается средняя величина.

Так, например, даные значения масс модифицированного волокна (табл.1) располагаем в ряд в порядке возрастания.

Первая группа образцов: 22,27; 22,36; 22,79; 23,22; 24,1 ;24,12; 24,22; 24,31.

Вторая группа образцов: 21,42; 21,62; 21,74; 22,82; 22,87; 23,08; 23,13; 23,25.

Третья группа образцов: 23,17; 22,47; 22,66; 22,92; 23,12; 23,2; 23,61; 23,72.

Значение медианы равно 23,660; 22,845; 23,145 по группам, соответственно.

В Microsoft Excel для этого нужно использовать функцию **МЕДИАНА,** которая рассчитывает медиану для заданной группы аргументов.

Второй способ определения общей числовой характеристики признака заключается в выяснении той его величины, которая встречается чаще всего.

Полученная этим путем типичная величина признака называется «модой». В Microsoft Excel есть подобная функция - функция **МОДА**.

Третий способ определения типичной величины признака состоит в вычислении так называемого «среднего арифметического».

В Excel функция **СРЗНАЧ** выполнит эту задачу и рассчитает «среднее арифметическое» значений, заданных в списке аргументов по формуле (1)

$$\overline{M} = \frac{\sum X_i}{n} \qquad (1)$$

В нашем примере значения «среднего арифметического» по каждой группе будут равны 23,424; 22,491; 23,109, соответственно.

Все эти способы определения типичной величины могут быть применены к получению общих статистических характеристик какого угодно другого свойства или признака.

Из трех указанных выше способов определения средней (*типичной*) величины варьирующего признака в основном используется «среднее арифметическое».

Для определения граничных значений группы показателей, как правило, вычисляют их минимальное и максимальное значения. В Microsoft Excel функции **МИН** и **МАКС** выполнят эти действия.

Определение только одной средней величины М варьирующего признака для исчерпывающей его характеристики все же не является вполне достаточным. При сравнении, например, массы отдельного образца со средней массой группы образцов, иногда бывает важно определить степень отличия параметра от этой «нормы».

Вполне объективный критерий такой оценки может быть найден только на основании учета самой величины варьирования значений данного признака. Вполне очевидно, что оценка разницы между значением какого-либо параметра и средним значением зависит от изменчивости самого этого параметра.

Исходя из этого средняя величина варьирования данного признака, это числовая его характеристика, которая могла бы служить объективной мерой для оценки отличия отдельного параметра от среднего (*типичного*) значения параметра всей группы объектов исследования.

Для этого существуют несколько способов вычисления средней величины варьирования [3,4,5].

Первый способ – определение среднего отклонения, второй – определение среднеквадратического отклонения.

К примеру возьмем из таблицы данные по первой группе образцов: *22,27; 22,36; 22,79; 23,22; 24,1 ;24,12; 24,22; 24,31* и определим среднее арифметическое М=*23,424*. Сравним теперь каждое значение с М и выясним, на сколько масса модифицированного волокна образцов первой группы отличается от среднего значения.

Получаем при этом: *-0,696; 0,634; -0,676; 0,204; 1,064; -0,886; -0,796; 1,154*.

Из этого примера видно, что значения отклонений могут быть как положительными, так и отрицательными. В данном случае это говорит о

том, что масса первого образца на 0,696 меньше среднего значения, а второго – на 0,634 больше и т.д.

Теперь среднее значение отклонений или просто «среднее отклонение» мы можем найти, сложив абсолютную величину всех отклонений и их сумму разделив на число этих отклонений. В нашем примере получим: 0,764.

Эти вычисления в Microsoft Excel выполняет функция **СРОТКЛ**, которая возвращает среднее абсолютных значений отклонений данных от среднего рассчитанное по формуле (2)[6].

$$СРОТКЛ = \frac{1}{n}\left|x - \bar{x}\right| \qquad (2)$$

Полученное значение (***в нашем примере 0,764***) может служить числовой характеристикой среднего отклонения массы отдельных образцов от средней массы образцов всей группы.

В самом деле, при малом размахе колебаний массы отдельных образцов, т.е. когда все измерения окажутся тесно сгруппированными около М, отдельные отклонения, общая их сумма и среднее отклонение также окажутся малыми; при большом же размахе колебаний отдельных измерений, т.-е. при сильной разбросанности соответствующих чисел, отдельные отклонения неизбежно окажутся большими, что тотчас же отразится и на увеличении среднего отклонения.

Поэтому, вычисленное этим способом среднее отклонение вполне точно характеризует величину изменчивости данного признака, увеличиваясь или уменьшаясь вместе с изменениями в степени его варьирования.

При вычислении среднего отклонения таким способом мы сталкиваемся с некоторым компромиссом. Дело в том, что, складывая отдельные отклонения для вычисления из них среднего отклонения, мы не имеем права отбрасывать знаки. Складывая же эти отклонения с учетом знаков, мы всегда получим в общем итоге ноль.

В выходе из этой ситуации поможет применение особого дополнительного приема, заключающегося в искусственном превращении всех отклонений (*как положительных, так и отрицательных*) только в одни положительные отклонения посредством временного возведения их в квадрат.

В нашем примере после возведения в квадрат получим следующий ряд значений: 0,485; 0,402; 0,457; 0,042; 1,132; 0,785; 0,634; 1,331, сумма которого будет равна 5,267, а среднее значение равно 0,658.

Полученное в нашем случае среднее значение 0,658 выражает собою не среднюю величину отклонения, а среднюю величину его квадрата.

Чтобы найти нужное нам среднее отклонение следует извлечь из полученного нами числа (0,658) квадратный корень. Следовательно, искомое среднее отклонение будет равно 0,811.

Вычисленное этим путем среднее отклонение называется «средним квадратическим отклонением» и имеет двойной знак (*плюс-минус*). Среднее квадратическое отклонение всегда оказывается несколько больше простого среднего отклонения. Это объясняется тем, что при возведении отдельных отклонений в квадрат особенно сильно возрастают большие отклонения, увеличивая собой общую сумму их квадратов, а стало быть, и конечный результат вычислений.

На практике для характеристики средней величины варьирования обычно пользуются средним квадратическим отклонением, которое показывает средний размах колебаний отдельных значений какого-либо признака и может служить мерой изменчивости или варьирования этого признака, являясь той границей, которая отделяет малые отклонения от отклонений значительных.

Описанный способ вычисления s относится, собственно, к тому случаю, когда число отдельных измерений чрезвычайно велико, например, когда отклонение определяется из нескольких сотен и даже из нескольких тысяч отдельных измерений [7,8].

В случае, когда число измерений незначительно, следует внести в этот способ некоторую поправку. Поправка заключается в том, что сумму квадратов отклонений перед извлечением корня делят не на число всех измерений, а на другое число, на единицу меньшее.

Конечная функция для расчета среднего квадратического отклонения для небольшого числа измерений в Microsoft Excel называется **СТАНДОТКЛОН**, которая рассчитывает результат по формуле (3).

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{(n-1)}} \qquad (3)$$

где x *— выборочное среднее* ***СРЗНАЧ****(число1,число2,…),* а n *— размер выборки.*

При расчете среднего арифметического *M* по указанным выше правилам для какой-либо однородной группы объектов исследования, нельзя быть вполне уверенным в том, что полученный результат совершенно точно характеризует величину этого признака у всех других подобных объектов.

На результаты статистического анализа всегда влияет тот или иной подбор объектов исследования, который называют выборкой.

Обоснование возможности оценки генеральных характеристик выборочными, как для больших, так и для малых выборок базируется на условии независимости и случайности выборок. И то, и другое обеспечивается различными способами организации выборок.

Независимость выборок достигается их повторностью, т.е. возвратом выбранных членов. Однако при испытаниях, связанных с порчей испытуемого образца (*например, испытания на прочность*) возврат становится невозможным. В этих случаях выборки бесповторные; и независимость может быть только при очень большом объеме N генеральной совокупности.

Случайность выборок может быть обеспечена различными способами, в основу каждого из них положена некоторая вероятностная схема. Каждому из таких способов соответствуют свои формулы для доверительных границ и доверительных объемов.

Различают такие основные типы выборок.

1. Простая случайная выборка основана на принципах независимости способа отбора от изучаемого признака и равной возможности для всех единиц совокупности быть включенной в выборку. Сюда относятся ряд методов полевого опыта, методика отбора навесок при анализе семян и т. п.

2. Систематическая выборка, для образования которой единицы наблюдения или учета выбирают по определенной системе, установленной специальными исследованиями. По такому принципу организуют выборки, например, при сортовом контроле (*апробации*), при учете урожая и других показателей пробными площадками и т. д.

3. Типическая выборка производится из совокупности, состоящей из групп, резко различающихся в каком-либо отношении. В этом случае характеристика совокупности слагается из частных характеристик отдельных групп с учетом их удельного веса. Такой метод составления выборок используют, например, в работе с гибридами, когда для характеристики каждого типа растений отбирают в его пределах по правилам случайной выборки какую-то часть представителей и на основании частных показателей получают представление о всем потомстве.

4. Двухстадийная выборка. Особенность такой выборки состоит в том, что сначала из совокупности отбирают какую-то часть объектов (по правилам случайной или систематической выборки), а затем из нее производят выборку «второго порядка», которую непосредственно анализируют. По принципу такой выборки производят, например, анализ качества семян; сначала отбирают пробы N из партии семян, а затем из нее выделяют навеску (*тоже выборка*) для определения всхожести, чистоты и т.д.

По такому же способу образуют выборки для анализа химического состава и других аналогичных исследований.

Способ образования выборки учитывают при статистической обработке выборочных совокупностей, в частности при определении степени варьирования опытных данных.

Таким образом, зная о некоторой неточности, связанной с различными способами организации выборок, нужно в каждом

отдельном случае определять также и величину той ошибки, которая была допущена при вычислении.

Определение средней ошибки представляет собой не что иное, как средний размах возможных колебаний нескольких средних арифметических значений, вычисленных для вполне однородного материала, но при различной его группировке. Для определения ошибки используется функция **СТОШYX** возвращающая стандартную ошибку предсказанных значений *y* для каждого значения *x* в регрессии. Уравнение для расчета стандартной ошибки имеет следующий вид (4)[6]:

$$k = \sqrt{\frac{1}{(n-2)}\left[\sum (y-\bar{y})^2 - \frac{\left[\sum (x-\bar{x})(y-\bar{y})\right]^2}{\sum (x-\bar{x})^2}\right]} \qquad (4)$$

где *x и y — выборочные средние значения* ***СРЗНАЧ*** *(известные значения - x) и* ***СРЗНАЧ*** *(известные значения - y), а n — размер выборки.*

В научных исследованиях при обработке каких-либо полученных результатов измерения часто возникает вопрос об установлении связи между двумя показателями в одной выборке. Для решения такой задачи используют корреляционный анализ (эффективный метод, который разрешает анализировать значительные объемы информации с целью исследования вероятной взаимосвязи двух или больше переменных).

Иными словами, корреляционный анализ помогает установить, можно ли предсказывать возможные значения одного показателя, зная величину другого.

Как правило, используется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде. Коэффициент корреляции выборки представляет собой ковариацию двух наборов данных, деленную на произведение их стандартных отклонений [9].

Корреляционный анализ дает возможность установить, ассоциированы ли наборы данных по величине, то есть, большие значения из одного набора данных связаны с большими значениями другого набора (*положительная корреляция*), или малые значения одного набора связаны с большими значениями другого (*отрицательная корреляция*), или данные двух диапазонов никак не связаны (*корреляция близка к нулю*).

В Excel функция **КОРРЕЛ** возвращает коэффициент корреляции между интервалами ячеек *массив 1* и *массив 2*.

Уравнение для расчета коэффициента корреляции имеет следующий вид (5).

$$r_{xy} = \frac{Cov(X,Y)}{s_x \cdot s_y} \qquad (5)$$

где x и y - выборочные средние значения ***СРЗНАЧ*** *(массив 1) и* ***СРЗНАЧ****(массив 2).*

Выводы

Использование современной вычислительной техники и программного обеспечения в ходе статистической обработки экспериментальных данных и оформления результатов исследования может оказать неоценимую помощь. Поэтому, рассмотрев возможности проведения статистического анализа данных с помощью одного из распространенных табличных процессоров – Microsoft Excel, можно прийти к такому выводу, что на начальном этапе он может вполне конкурировать с дорогим специализированным программным обеспечением, поскольку имеет необходимый набор функций и тем самым поможет исследователю, как обработать данные, так и подготовить результаты к графическому представлению и печати.

1. *Нестеренко Л.В.* Теоретичне обгрунтування модифікації лляного волокна / Л.В.Нестеренко, Т.Я.Тулученко, Л.А.Чурсіна // Легка промисловість. – 2003. – №4. – С.59.

2. *Нестеренко Л.В.* Визначення відсотка виходу модифікованого волокна з низькосортної лляної сировини / Л.В.Нестеренко, Т.Я.Тулученко, Л.А.Чурсіна, М.М.Кобельчук // Легка промисловість. – 2003. – №2. – С.58.

3. *Поморский Ю.Л.* Вариационная статистика /Ю.Л.Поморский.–1931. – 310с.

4. *Елисеева И.И.* Общая теория статистики /И.И.Елисеева, Юзбашев М.М.– М.: Финансы и статистика, 1995. – 386с.

5. *Лукьянова Н.Ю.* Статистический анализ данных с использованием компьютера /Н.Ю.Лукьянова. – Калининград: КГУ, 2001. – 89с.

6. *Макрова Н.В.* Статистика в Excel /Н.В.Макрова, В.Я.Трофимец.– М.: Финансы и статистика, 2002. – 361с.

7. *Виноградов Ю.С.* Математическая статистика и ее применение к исследованиям в текстильной промышленности /Ю.С.Виноградов. – М.: Легкая индустрия, 1964.– 320с.

8. *Вольф В.Г.* Статистическая обработка опытных данных /В.Г.Вольф. – М.: Колос, 1966. – 254с.

9. *Андреев С.А.* Программы для статистической обработки результатов сельскохозяйственного эксперимента на программируемых микрокальк-куляторах /С.А.Андреев. – М.: 1990. – 84 с.