И. А. Буяковская

ОБРАБОТКА СТАТИСТИЧЕСКИХ ДАННЫХ НА ОСНОВЕ ПРОГРАММ ДЛЯ ОС LINUX

В данной статье приводится обзор программных средств, входящих в состав ПСПО и реализующих статистический анализ данных.

Решение задач математической статистики связано как с расчетом количественных показателей, так и с графической интерпретацией полученных результатов.

Достаточно удобным инструментом для анализа статистических данных являются электронные таблицы: OpenOffice.org Calc, Gnumeric. В них можно выполнять операции сортировки и ранжирования данных, вычисления описательных статистик, построения некоторых видов графиков, представления распределения дискретной случайной величины и интервальных данных. Причем эти операции выполнимы как на основе использования непосредственно функций из категории статистические, так и из других категорий (логические, электронная таблица, математические), а также на основе надстройки «Подбор параметра».

Например рассмотрим следующую задачу, решение которой производится в программе OpenOffice.org Calc: В ряду чисел: 2, 7, 10, __, 18, 19, 27 одно число оказалось стертым. Восстановите его, зная, что среднее арифметическое этих чисел равно 14.

Воспользуемся надстройкой «Подбор параметра» для нахождения стертого элемента выборки . Для этого наберите элементы ряда в столбец А. В ячейку А9 внесите формулу нахождения среднего арифметического считая, что утерянный элемент расположен в ячейке А7. Таким образом, формула будет выглядеть следующим образом: =SUM(A1:A7)/7. Для запуска надстройки выбираем команду: *Сервис - Подбор параметра.* В открывшемся диалоговом окне задаем следующие настройки:

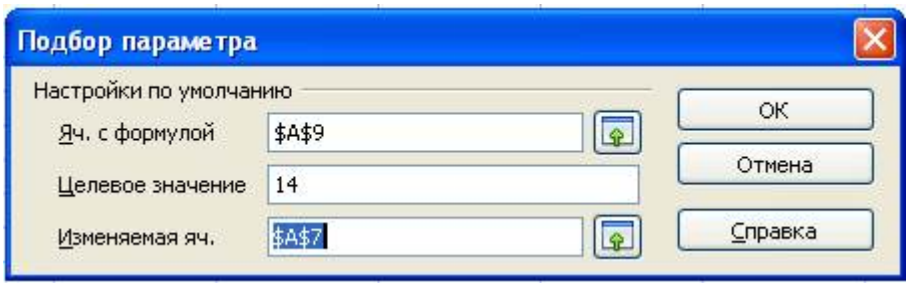

Рис. 1. Диалоговое окно «Подбор параметра»

В результате в ячейку А7 будет получено значение утерянного элемента ряда.

	A
1	2
2	7
3	10
4	18
5	19
6	27
7	15
8	среднее
9	14

Рис. 2. Результат вычисления на основе надстройки «Подбор параметра»

Решим задачу на графическую интерпретацию статистических данных: Проверочную работу по алгебре выполняли 180 учащихся. В результате группировки работ по полученным оценкам составили таблицу:

Оценка	1	2	3	4	5
Частота	0	16	77	65	22

На основании данных этой таблицы постройте полигон распределения оценок за проверочную работу.

Для построения полигона отмечают в координатной плоскости точки, абсциссами которых служат оценки, а ординатами – соответствующие им частоты.

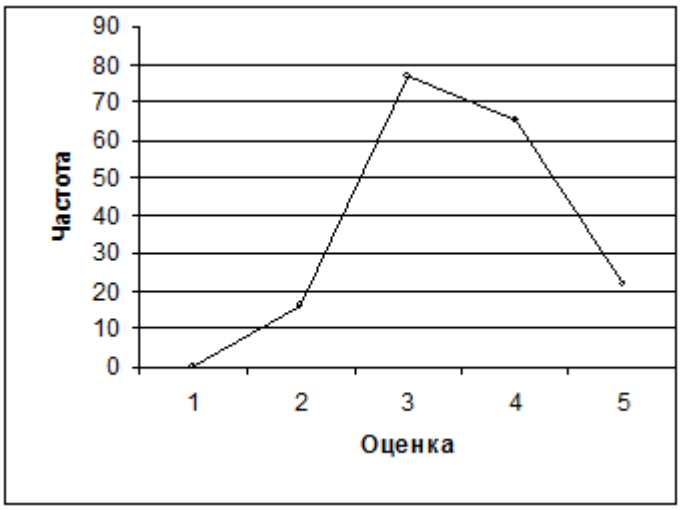

Рис. 3. Полигон распределения оценок

Но полноценная обработка результатов исследования должна выполняться в специализированных пакетах. Универсальным пакетом из свободного программного обеспечения является PSPP, который часто сопоставляют с лицензионным пакетом SPSS предназначенного для статистического анализа данных. В пакете PSPP предусмотрен рассчет основных числовых характеристик входящих в состав описательной статистики, решаются задачи линейной регрессии, включены такие статистические методы анализа как: t-тест и непараметрические критерии. Он разработан для выполнения анализа входных данных как можно быстрее, независимо от их размера.

Также в качестве одного из направлений рассматривается применение математических пакетов позволяющих выполнять вычисления и построение статистических диаграмм и графиков. В данном случае можно привести такие как: Scilab, Maxima и Mathomatic.

В то же время применимо использование таких узкоспециализированных пакетов предназначенных только для анализа и визуализации научных данных как: KmPlot, QtiPlot, Gnuplot, UDAV, MayaVi, Open Data Explorer (OpenDX). Продукт Open Data Explorer является версией пакета визуализации данных и разработки приложений корпорации IBM с открытым исходным кодом, который необходим для выполнения крайне сложных задач визуализации научных данных. Программа MayaVi представляет собой инструмент визуализации данных, связываемый с языком Python.