

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

1. Предварительная обработка данных

Основным объектом исследования в математической статистике является выборка. Выборкой объема n называются числа $x_1, x_2, \dots, x_n$, получаемые на практике при n – кратном повторении эксперимента в неизменных условиях. На практике выборку чаще всего представляют статистическим рядом. Для этого вся числовая ось, на которой лежат значения выборки, разбивается на k интервалов (это число выбирается произвольно от 5 до 10), которые обычно равны, вычисляются середины интервалов z_i , и считается число элементов выборки, попадающих в каждый интервал n_i . Статистическим рядом называется последовательность пар (z_i, n_i) . Рассмотрим решение задачи на ЭВМ в программе EXCEL на следующем примере.

ПРИМЕР 1. Дана выборка числа проданных автомобилей торговой фирмой за 25 недель:

14, 18, 16, 21, 12, 19, 27, 19, 15, 20, 27, 29, 22, 28, 19, 17, 18, 24, 23, 22, 19, 20, 23, 21, 19.

Построим статистический ряд, полигон, гистограмму и кумулятивную кривую. Откроем книгу программы EXCEL, Введем в первый столбец (ячейки A1-A25) исходные данные. Определим область чисел, на какой лежат данные. Для этого найдем максимальный и минимальный элементы выборки. Введем в B1 подпись «Максимум», а в B2 - подпись «Минимум». В соседних ячейках C1 и C2 определим функции «MAX» и «MIN». Для этого ставим курсор в C1 и вызываем мастер функций, нажав на кнопку , в открывшемся окне в поле «Категория» выбираем «Статистические», и ниже ищем функцию МАКС и вызываем ее двойным щелчком мыши по названию. В качестве аргумента функции (в графе «Число 1») обведем область данных (ячейки A1-A25). Поле «Число 2» оставляем пустым. Нажимаем «ОК». Результатом будет число 29. Ставим курсор в ячейку C2 и аналогично вводим функцию МИН. Результат – число 12. Видно, что все данные укладываются на отрезке [12;30]. Разделим его на девять (выбирается произвольно от 5 до 10) интервалов по 2 единицы каждый:

12-14, 14-16, 16-18, 18-20, 20-22, 22-24, 24-26, 26-28, 28-30. В ячейки D1-D9 вводим верхние границы интервалов группировки – числа 14, 16, 18, 20, 22, 24, 26, 28, 30. Для вычисления частот n_i используют функцию ЧАСТОТА, находящуюся в категории «Статистические». Введем ее в ячейку E1. В строке «Массив данных» введем диапазон выборки (ячейки A1-A25). В строке «Массив интервалов» введем диапазон верхних границ интервалов группировки (ячейки D1-D9). Результат функции является массивом и выводится в ячейках E1-E9. Для полного вывода (не только первого числа в E1) нужно выделить ячейки E1-E9, обведя их мышью, и нажать F2, а далее одновременно CTRL+SHIFT+ENTER. Результат – частоты интервалов 2,2,3,7,4,3,0,3,1.

Для построения гистограммы нужно выбрать ВСТАВКА/ДИАГРАММА или нажать на соответствующий значок на основной панели (при этом курсор должен стоять в свободной ячейке). Далее выбрать тип: ГИСТОГРАММА, вид по выбору, нажать «ДАЛЕЕ», в строке «ДИАПАЗОН» обвести частоты E1-E9, перейти на вкладку «РЯД», в строке «ПОДПИСИ ОСИ X» ввести интервалы в ячейках D1-D9,

нажать «ДАЛЕЕ» ввести название «ГИСТОГРАММА», подписи осей: ось X - «ИНТЕРВАЛЫ» и ось Y - «ЧАСТОТА», нажать «ГОТОВО». Для создания полигона перейти на пустую ячейку и сделать то же самое, только вместо типа диаграммы «ГИСТОГРАММА», выбрать «ГРАФИК». Для построения кумулятивной кривой нужно посчитать накопленные частоты. Для этого в ячейку F1 вводим «=E1», в F2 – вводим «=F1+E2» и автозаполнением перетаскиваем эту ячейку до F9. Далее строим график как и в случае полигона, но в строке «ДИАПАЗОН» вводим накопленные частоты, ссылаясь на F1-F9, а на вкладке «РЯД», в строке «ПОДПИСИ ОСИ X» вводим интервалы в ячейках D1-D9.

Задание 1. Дана выборка выручки магазина за последние 30 дней. Составить статистический ряд, построить гистограмму, полигон, кумуляту.

Выборка														
18	19	21	18	16	19	18	16	17	18	15	22	18	17	22
14	19	16	14	14	22	14	21	18	16	12	19	18	18	15

2. Точечное оценивание

Для исследования основных свойств явления или объекта, представленного выборкой вычисляют точечные и интервальные оценки.

Точечные оценки параметров распределения это оценки, полученные по выборке и приближенно равные оцениваемым параметрам. Основными точечными оценками являются:

Объем выборки n – количество элементов в выборке.

Выборочное среднее $\bar{x}$ – оценка математического ожидания, среднеарифметическое элементов выборки.

Выборочная дисперсия S^2 – среднее квадратов отклонения элементов выборки от выборочного среднего, является оценкой дисперсии, характеризует разброс выборочных значений.

Стандартное отклонение S – корень из дисперсии.

Медиана h – средний элемент вариационного ряда или полусумма двух средних элементов, если объем выборки четный.

Мода d – наиболее часто повторяющийся элемент.

Коэффициент эксцесса δ - характеризует «островерхность» гистограммы или полигона по сравнению с кривой Гаусса нормального распределения.

Коэффициент асимметрии γ - характеризует степень симметричности гистограммы или полигона.

Перцентиль на уровне p - значение t_p , меньше которого $p \cdot 100\%$ элементов выборки.

ПРИМЕР 2. Имеется выборка числа автомобилей, проданных автосалоном за 25 недель: 43, 38, 34, 51, 47, 45, 41, 52, 50, 38, 43, 44, 39, 46, 49, 42, 42, 38, 53, 55, 48, 45, 41, 49, 47. Найти основные числовые характеристики выборки.

Запускаем программу EXCEL, первый лист. Вводим исходные данные в ячейки A1-A25. Находим числовые характеристики. Для ввода функций выделяем два столбца, например B и C, в первом вводим название характеристики, во втором – функцию. В ячейки B1-B11 вводим подписи числовых характеристик, то есть

вписываем в эти ячейки первый столбец таблицы приведенной ниже. В C1 вводим текст «Функция» и ниже определяем функции, соответствующие названию (из второй колонки таблицы). Все функции вызываются нажатием на кнопку , находятся в категории «Статистические» и в качестве массива данных (поле «ЧИСЛО 1»), указывается ссылка на A1-A25. Например, для ввода первой из них ставим курсор в C2, нажимаем , выбираем категорию «Статистические» и функцию «Счет», в открывшемся окне ставим курсор в поле «Число 1» и обводим курсором ячейки A1-A25, нажимаем «ОК». Также поступаем и с другими функциями.

Характеристика	Функция
Объем выборки	СЧЁТ(массив данных)
Выборочное среднее	СРЗНАЧ(массив данных)
Дисперсия	ДИСП(массив данных)
Стандартное отклонение	СТАНДОТКЛОН(массив данных)
Медиана	МЕДИАНА(массив данных)
Мода	МОДА(массив данных)
Коэффициент эксцесса	ЭКСЦЕСС(массив данных)
Коэффициент асимметрии	СКОС(массив данных)
Перцентиль 40%	ПЕРСЕНТИЛЬ(массив данных; 0,4)
Перцентиль 80%	ПЕРСЕНТИЛЬ(массив данных; 0,8)

Существует другой способ вычисления числовых характеристик выборки. Для этого ставим курсор в свободную ячейку (например, D1). Затем вызываем в меню «Данные» подменю «Анализ данных». Если в меню «Данные» отсутствует этот пункт, то в обратитель к преподавателю для его подключения. В окне «Анализ данных» нужно выбрать пункт «Описательная статистика». В появившемся окне в поле «Входной интервал» делаем ссылку на выборку A1-A25, помещая курсор в поле и обводя эти ячейки. Оставляем группирование «По столбцам». В разделе «Параметры вывода» ставим флажок на «Выходной интервал» и в соседнем поле задаем ссылку на верхнюю левую ячейку области вывода (например, D1), ставим флажок напротив «Описательная статистика», нажимаем «ОК». Результат – основные характеристики выборки (сделайте шире столбец D, переместив его границу в заголовке).

Задание 2. Для данных из задания 1 вычислить основные числовые характеристики выборки обоими способами.

3. Интервальное оценивание.

Рассмотрим теперь методы интервального оценивания. Доверительным интервалом называется интервал (a;b), в который с заданной вероятностью p попадает оцениваемый параметр. Вероятность p называется доверительной. Вместо нее часто задают величину $\alpha=1-p$, называемую уровнем значимости. Если выборка объема n представляет случайную величину, распределенную нормально, то доверительные интервалы для матожидания и дисперсии равны

$$m \in \left(\bar{x} - \frac{S \cdot t_{1-\alpha/2}(n-1)}{\sqrt{n}}; \bar{x} + \frac{S \cdot t_{1-\alpha/2}(n-1)}{\sqrt{n}} \right),$$
$$\sigma^2 \in \left(\frac{S^2 \cdot (n-1)}{\chi^2_{1-\alpha/2}(n-1)}; \frac{S^2 \cdot (n-1)}{\chi^2_{\alpha/2}(n-1)} \right),$$

где $t_p(n)$ и $\chi^2_p(n)$ - квантили распределения Стьюдента и хи-квадрат, $\alpha=1-p$.

Возвращаемся на лист 1 электронной таблицы с данными примера и для них вычислим доверительные интервалы при p=0,05. Вводим данные согласно рисунку:

	F	G	H	I	
1		Уровень значимости		0,05	
2		Интервал	Левая граница	Правая граница	
3		Матожидание			
4		Дисперсия			

Для вычисления величины $\frac{S \cdot t_{1-\alpha/2}(n-1)}{\sqrt{n}}$ служит функция «ДОВЕРИТ» категории

«Статистические» с тремя параметрами «Альфа» - уровень значимости $\alpha=1-p$, «Станд_откл» - среднеквадратическое отклонение S, «Размер» - объем выборки n. Таким образом, вводим в H3 функцию:

=СРЗНАЧ(A1:A25)-ДОВЕРИТ(I1;СТАНДОТКЛОН(A1:A25);25)

а в ячейку I3 функцию:

=СРЗНАЧ(A1:A25)+ДОВЕРИТ(I1;СТАНДОТКЛОН(A1:A25);25)

Для вычисления доверительного интервала для дисперсии следует отметить, что функция вычисления квантили распределения хи-квадрат (обратного распределения хи-квадрат) называется «ХИ2ОБР» (категория «Статистические») и имеет два параметра: первый «Вероятность» содержит доверительную вероятность p, второй – степень свободы n-1. Вводим в соответствии с данными условиями и формулой для доверительного интервала в ячейку H4 запись:

=ДИСП(A1:A25)*24/ХИ2ОБР(0,025;24)

а в ячейку I4 запись: =ДИСП(A1:A25)*24/ХИ2ОБР(0,975;24).

Получаем значения границ доверительных интервалов.

Задание 3. Для данных из задания 1 вычислить доверительные интервалы для матожидания и дисперсии при $\alpha=0,01$. Изменяя значения уровня значимости α сделать вывод о его влиянии на ширину интервала.