ЛАБОРАТОРНАЯ РАБОТА № 1
СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ

Цель: Научиться основным методам обработки данных, представленных выборкой, путем построения гистограммы, определения выборочного среднего, выборочной дисперсии, выборочной медианы и моды.

Вероятностная модель ставит в соответствие результатам наблюдений

$$x_1, x_2, \dots, x_n \tag{1}$$

последовательность случайных величин

$$X_1, X_2, \dots, X_n. \tag{2}$$

Предполагается, что случайные величины $X_1, X_2, \dots, X_n$ независимы и имеют одно и то же распределение $F(x)$. Полагают, что наблюдения (1) являются значениями величин (2) при осуществлении вероятностного эксперимента. Несмотря на различие объектов (1) и (2), в математической статистике принято называть и то и другое **выборкой из генеральной совокупности**.

Количество наблюдений n называется объемом выборки.

Произвольная случайная величина X характеризуется своей функцией распределения вероятностей $F(x)$. Если эта функция неизвестна, но известна выборка (1), числовые данные которой являются значениями случайной величины X, то возможно построить **эмпирическую функцию распределения вероятностей** $F_n(x)$, которая служит оценкой теоретической функции распределения вероятностей $F(x)$. Если обозначить через $\mu_n(x)$ число тех значений $x_1, x_2, \dots, x_n$, которые меньше или равны x, то

$$F_n(x) = \frac{\mu_n(x)}{n}. \tag{3}$$

Если объем выборки n большой, то для представления о виде ее распределения строится **гистограмма**.

Вводим в первый столбец (ячейки A1…) исходные данные. Для элементов выборки находим минимальный и максимальный элементы,

которые ограничивают интервал, содержащий все элементы выборки. Для этого запишем в первую строку второго столбца (B1) слово Максимум, а во вторую строку второго столбца (B2) слово 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 fx , в открывшемся окне в поле «Категория» выбираем **СТАТИСТИЧЕСКИЕ**, и ниже ищем функцию МАКС и вызываем ее двойным щелчком по названию. В качестве аргумента функции (в графе «Число 1») обведем область данных (ячейки A1…). Поле «Число 2» оставляем пустым. Нажимаем «ОК». Ставим курсор в ячейку C2 и аналогично вводим функцию **МИН**. В некоторых случаях для удобства обработки интервал расширяется, но не существенно.

Следующим шагом является разбиение построенного интервала на 5-10 более мелких интервалов. Если разбиение построено удачно, то гистограмма будет напоминать график плотности (если она существует) распределения вероятностей случайной величины, значениями которой являются элементы выборки. Если разбиение мелкое, то гистограмма не дает представления о плотности распределения вероятностей из-за случайных флуктуаций. Если разбиение крупное, то гистограмма также не дает представления о плотности распределения вероятностей из-за того, что теряется много информации.

Чтобы построить интервалы разбиения (группировки), нужно от максимального значения выборки вычесть минимальное значение и полученный результат разделить на число интервалов. Полученное значение называется шагом разбиения. Чтобы получить верхние границы интервалов группировки, нужно последовательно прибавлять шаг разбиения, начиная от минимального значения выборки.

В ячейки D1… вводим верхние границы интервалов группировки. Для вычисления частот n_i используется функция **ЧАСТОТА**, находящаяся в категории **СТАТИСТИЧЕСКИЕ**. Введем ее в ячейку E1. В строке «Массив данных» введем диапазон выборки (ячейки A1…). В строке «Массив интервалов» введем диапазон верхних границ интервалов группировки

(ячейки D1…). Результат функции является массивом и выводится в ячейках E1... Для полного вывода (не только первого числа в E1) нужно выделить ячейки E1…, обведя их мышью, и нажать F2, а далее одновременно CTRL+SHIFT+ENTER. Результат – частоты n_i, которые показывают, сколько элементов выборки попало в каждый из интервалов разбиения.

Для построения гистограммы в EXCEL 2003 нужно из меню **ВСТАВКА** выбрать **ДИАГРАММА** (или нажать на соответствующий значок **МАСТЕР ДИАГРАММ** на основной панели), при этом курсор должен стоять в свободной ячейке. Далее выбрать тип диаграммы: **ГИСТОГРАММА**, вид по выбору, нажать **ДАЛЕЕ**, в строке **ДИАПАЗОН** обвести частоты E1.., перейти на вкладку **РЯД**, в строке **ПОДПИСИ ОСИ** X ввести интервалы в ячейках D1…, 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**.

При использовании EXCEL 2007 для создания диаграммы необходимо выделить блок данных, на основании которых строится диаграмма. В выделяемый блок данных включить не только числовые данные, но и заголовки строк (столбцов), в которых они расположены. Заголовки будут использованы в качестве подписей по осям (меток) и для формирования условных обозначений (легенды). При выделении блоков с данными для построения диаграмм необходимо соблюдать два правила:

1. Выделенный фрагмент должен состоять из равновеликих столбцов.
2. В выделенном фрагменте не должно быть объединенных ячеек.

Для построения гистограммы необходимо перейти на вкладку **ВСТАВКА**, открыть список **ГИСТОГРАММА** выбрать нужную гистограмму. Гистограмма строится сразу. Иногда необходимо выделить построенную диаграмму и провести изменение размера шрифта или растянуть диаграмму для лучшего чтения данных в поле диаграммы. Если вызвать контекстное меню в поле всей диаграммы, то меню предлагает три отдельных шага в

построении диаграммы (в предыдущих версиях было четыре шага): Изменить тип диаграммы; выбрать данные; переместить диаграмму.

В мастере функций fx существуют специальные функции, позволяющие вычислять выборочные характеристики.

Функция **СРЗНАЧ** вычисляет выборочное среднее (оценку теоретического математического ожидания)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Функция **ДИСП** вычисляет выборочную дисперсию (оценку теоретической дисперсии)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Функция **СТАНДОТКЛОН** вычисляет квадратный корень из выборочной дисперсии.

Функция **МЕДИАНА** вычисляет выборочную медиану (оценку медианы) заданной выборки. Медианой случайной величины называется то ее значение, которое делит распределение на две равновероятные половины. В качестве выборочной медианы m в выборке объема $2n+1$ берут значение x_{n+1} в вариационном ряде. Если объем выборки равен $2n$, то в качестве выборочной медианы m берут $\frac{1}{2}(x_n + x_{n+1})$.

Функция **МОДА** вычисляет выборочную моду (оценку моды). Модой случайной величины называется ее наиболее вероятное значение.

В Excel можно генерировать случайные числа, имеющие разные законы распределения. Для этого можно использовать надстройку **АНАЛИЗ ДАННЫХ** и пункт **ГЕНЕРАЦИЯ СЛУЧАЙНЫХ ЧИСЕЛ**.

Если вы хотите сгенерировать, например, 100 случайных чисел из нормального распределения, то в поле **ЧИСЛО ПЕРЕМЕННЫХ** введите 1; в поле **ЧИСЛО СЛУЧАЙНЫХ ЧИСЕЛ** введите 100; в списке **РАСПРЕДЕЛЕНИЕ** выберите **НОРМАЛЬНОЕ**; введите параметры

нормального распределения – **СРЕДНЕЕ** и **СТАНДАРТНОЕ ОТКЛОНЕНИЕ**. В качестве типа распределения можно выбрать, например, **РАВНОМЕРНОЕ, БИНОМИАЛЬНОЕ** или **РАСПРЕДЕЛЕНИЕ ПУАССОНА**. Введя для каждого распределения соответствующие параметры, получим сгенерированные случайные числа.

Задание

Сгенерировать 100 значений некоторой случайной величины, имеющей нормальное распределение. Построить гистограмму, вычислить выборочное среднее, выборочную дисперсию (исправленную), выборочные медиану и моду.