2. ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Статистическими данными называют данные, связанные с массовыми явлениями (как правило, эти данные являются неполными). Предметом математической статистики является разработка методов систематизации и анализа статистических данных. В число таких методов входят методы построения вероятностных (статистических) моделей явлений – моделей, основанных как на априорных предположениях об объективных закономерностях, порождающих данные, так и на эмпирических результатах регистрации данных.

С использованием статистических моделей решаются важнейшие задачи прикладной статистики – проверка статистических гипотез и нахождение статистических оценок. *Статистическими оценками* называют функции от наблюдаемых значений. Скалярные оценки называют *точечными*. К точечным оценкам предъявляются требования:

– *несмещенности*: математическое ожидание оценки должно быть равно оцениваемой величине;

– *эффективности*: среднеквадратичная ошибка оценки должна быть наименьшей среди всех возможных оценок;

– *состоятельности*: оценка должна сходиться по вероятности к оцениваемому параметру (при увеличении объема выборки вероятность сколь угодно малого отклонения оценки от оцениваемого параметра должна стремиться к нулю).

Если статистическая модель включает конечномерный вектор параметров (определяемых априорно и/или на основе эксперимента), то такую модель называют *параметрической*. Примером параметрической вероятностной модели является экспериментально-статистическая регрессионная модель (п. 5).

Описательная (*дескриптивная*) *статистика* – это совокупность методов обработки данных, не включающая построение параметрических моделей. Центральное положение в описательной статистике занимают методы первичного анализа выборки значений одного признака.

Генеральной совокупностью будем называть произвольное числовое множество X, а *выборкой* – любое его подмножество $\{x_i\} \subset X$. Элементы выборки называют *вариантами*, а их полное число n (с учетом повторений) называют *объемом выборки*. *Частотой* варианты x_i называют число n_i ее вхождений в выборку. *Относительной частотой* называют частное от деления частоты на объем выборки. Если все частоты вариант равны единице, то выборку называют *бесповторной*.

Последовательность вариант и соответствующих им частот, упорядоченная по возрастанию, называется *дискретным вариационным рядом*. После его построения для бесповторной выборки оценку медианы генеральной совокупности можно найти как значение, соответствующее «центру» ряда:

$$Me = \begin{cases} x_{(n+1)/2}, & n = 2k+1 \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}, & n = 2k \end{cases}, \; k \in N .$$

Оценкой математического ожидания является *выборочное среднее* – среднее арифметическое вариант:

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

Для характеристики «рассеяния» значений около «центра» используют оценки дисперсии, среднего квадратичного и среднего абсолютного отклонения.

Несмещенная *оценка дисперсии* вычисляется по формуле:

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

Если из каких-либо посторонних соображений для генеральной совокупности уже известно истинное математическое ожидание $M[X]$, то несмещенной оценкой дисперсии будет величина $\frac{1}{n}\sum_{i=1}^{n}(x_i - M[X])^2$.

Оценка стандартного отклонения связана с оценкой дисперсии:

$$s = \sqrt{s^2} ;$$

эта оценка не является несмещенной, но на практике используют именно ее по причине простоты отыскания. Если генеральная совокупность подчинена нормальному закону, то можно найти и несмещенную оценку стандартного отклонения; для выборки объема $n = 10$ она отличается (в большую сторону) от $\sqrt{s^2}$ на 3%, для выборки объема $n = 1000$ отличие составляет менее 0,03%.

В теории вероятностей термины «стандартное отклонение» и «среднее квадратичное отклонение» равноправны. В статистической литературе их часто применяют к оценкам и могут наделять различным смыслом; например, термин «стандартное отклонение» используют для обозначения величины

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} ,$$

в то время как термином «среднее квадратичное отклонение» некорректно обозначают величину

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} ,$$

или наоборот.

Следует всюду перед статистическими аналогами показателей распределений (генеральных совокупностей) использовать слово «оценка» (или добавлять перед названием величины слово «выборочная»). Числовая характеристика вероятностного распределения ни в коем случае не тождественна близкой к ней величине, вычисленной на основе опытных данных. Это подчеркивается и пра-

вилами записи; так, оценка стандартного отклонения обозначается буквой s, в то время как само стандартное отклонение обозначается буквой σ.

Стандартная ошибка оценки математического ожидания вычисляется как частное от деления оценки стандартного отклонения на корень из объема выборки:

$$s_{err} = \sqrt{\frac{s^2}{n}}.$$

Безразмерный *коэффициент вариации* вычисляется как частное от деления оценок стандартного отклонения и математического ожидания:

$$v = \frac{\sqrt{s^2}}{\bar{x}}.$$

Характеристиками рассеяния также являются *нижняя* $x_{1/4}$ и *верхняя* $x_{3/4}$ *квартили* – вычисляемые по выборке (т.е. являющиеся оценками соответствующих квантилей) процентные точки, для которых числа вариант, удовлетворяющих неравенствам $x_i < x_{1/4}$ и $x_i < x_{3/4}$, составляют 25% и 75%, соответственно.

Оценки моментов третьего и четвертого порядков и связанные с ними безразмерные оценки асимметрии и эксцесса без необходимости использовать не следует.

Для выборки большого объема дискретный вариационный ряд теряет наглядность. Принято выполнять группировку данных, разбивая весь диапазон

$$\left[x_{\min} \le \min\{x_i\}; x_{\max} \ge \max\{x_i\}\right]$$

изменения исследуемого признака (диапазон, включающий минимальное и максимальное значение вариант) на l частичных интервалов – разрядов, число которых выбирают по *правилу Стерджеса*:

$$l = 1 + 3{,}3 \lg n,$$

где n – объем выборки. При этом длины разрядов обычно равны между собой:

$$\Delta x = x_j - x_{j-1} = \frac{x_{\max} - x_{\min}}{l},\ j = \overline{1,l},\ x_0 = x_{\min},\ x_l = x_{\max},$$

а границы разрядов находятся в точках

$$x_k = x_{\min} + k\Delta x,\ k = \overline{1,l-1}.$$

Частоты n_j, соответствующие каждому разряду, находятся как суммы частот всех вариант, попавших в этот разряд. Для бесповторной выборки частота равна числу попавших в разряд вариант. Относительной частотой разряда называют частное n_j/n от деления частоты разряда на объем выборки.

Графическим представлением непрерывного вариационного ряда является *гистограмма* – ступенчатая фигура, состоящая из прямоугольников, основания которых построены на соответствующих разрядах, а высоты h_j равны частным от деления относительных частот на длины разрядов:

$$h_j = \frac{n_j}{n\Delta x}.$$

Гистограмма позволяет сделать предварительное суждение о плотности распределении генеральной совокупности. По гистограмме обычно находят оценку моды. Для этого на гистограмме находят прямоугольник с наибольшей высотой и проводят из противоположных вершин его верхнего основания два отрезка к противоположным вершинам верхних оснований соседних прямоугольников. В качестве оценки моды принимается абсцисса точки пересечения отрезков.