Краткие сведения из теории

Функция распределения F(x) случайной величины X обычно неизвестна, в лучшем случае она известна с точностью до некоторых неизвестных параметров. Сведения о распределении случайной величины и ее характеристиках можно получить, если имеются независимые многократные повторения опыта, в котором измеряются значения интересующей нас случайной величины.

Предположим, что независимые наблюдения над случайной величиной X позволили получить выборочную совокупность $x_1,x_2,\ldots,x_n$ объема n.

Выборочной (эмпирической) функцией распределения случайной величины X, построенной по выборке $x_1,x_2,\ldots,x_n$, называется функция $F_n(x)$, равная доле таких значений x_i, что $x_i \le x$, i=1,2,...,n. Иначе говоря, $F_n(x)$ есть частота события ($x_i \le x$) в выборке $x_1,x_2,\ldots,x_n$. Таким образом, эмпирическую функцию выборки $x_1,x_2,\ldots,x_n$ можно рассматривать как функцию распределения вероятностей, где каждому значению x_i, i=1,2,...,n приписана вероятность $\frac{1}{n}$. Установлено, что с ростом объема выборки n эмпирическая функция распределения равномерно по x приближается к функции распределения случайной величины X, то есть $\max_x |F_n(x) - F(x)| \to 0$ при $n \to \infty$ с вероятностью 1.

Кроме эмпирической функции распределения случайную величину X характеризуют также и численные показатели, которые можно разбить на три группы:

- характеристики положения,
- характеристики разброса,
- характеристики асимметрии и эксцесса.

К *характеристикам положения* относятся:

- выборочное среднее (среднее арифметическое) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

- выборочная медиана - корень уравнения $F_n(x)=0.5$,

- выборочная мода - наиболее часто встречающееся значение в выборке,

- выборочное среднее геометрическое значение $\sqrt[n]{x_1 x_2 \ldots x_n}$.

Эти характеристики показывают, где находится центр группирования статистических данных.

К *характеристикам разброса* относятся:

- выборочная дисперсия (правильнее говорить, несмещенная оценка выборочной дисперсии) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$,

- выборочное среднее квадратическое отклонение $s = \sqrt{s^2}$,

- коэффициент вариации $v = \frac{s}{\bar{x}} \cdot 100\%$,

- минимальное значение - наименьшее значение в выборке x_{min},

- максимальное значение - наибольшее значение в выборке x_{max},

- размах - разность между максимальным и минимальным значениями $R=x_{max}-x_{min}$,

- нижняя квартиль - корень уравнения $F_n(x)=0.25$,

- верхняя квартиль - корень уравнения $F_n(x)=0.75$,

- межквартильный размах - разность между верхним и нижним квартилями.

Дисперсия, среднее квадратическое отклонение, размах и квартили характеризуют разброс или степень рассеяния статистических данных.

Коэффициент асимметрии определяет насколько асимметрично распределение данных. Положительные значения асимметрии говорят о том, что кривая распределения смещена вправо от среднего значения выборки, а отрицательная - влево.

Коэффициент эксцесса показывает однородным или неоднородным является распределение данных по отношению к нормальному (гауссову) распределению. Для нормального распределения коэффициент эксцесса равен нулю. Если коэффициент эксцесса отрицательный, то кривая будет плоской с короткими «хвостами». Если коэффициент положительный, то кривая либо очень крутая в центре, либо имеет довольно длинные «хвосты».