Основы математической статистики

Математическое ожидание *M(x)* есть то значение, к которому стремится среднее арифметическое значение $\bar{x}$ когда число испытаний *N* стремится к бесконечности

$$M(x) = \lim_{N\to\infty} \bar{x},$$

где $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

На практике, при ограниченном числе испытаний, $M(x) \approx \bar{x}$.

Таким образом, математическое ожидание характеризует *среднее значение случайной величины*.

Для дискретной случайной величины $M(x)$ находится по формуле

$$M(x)=\sum_{i=1}^{n} x_i \cdot p_i\ ,$$

где n – число различных значений x_i; p_i – вероятность этих значений, вычисляемая как относительная частота $P_i=\frac{m_i}{N}$.

Дисперсия ($D(x)$) (от лат. *dispersion* – рассеяние) характеризует средний разброс (рассеяние) значений случайной величины относительно математического ожидания.

Для дискретной случайной величины $D(x)$ находится по формуле

$$D(x) = \sum_{i=1}^{n} (\Delta x_i)^2 \cdot p_i$$

где Δx_i – абсолютное отклонение $\Delta x_i = x_i - M(x)$

$$D(x) = \sum_{i=1}^{n} (x_i - M(x))^2 \cdot p_i$$

Среднее квадратическое отклонение (σ) характеризует средний разброс (рассеяние) значений случайной величины относительно математического ожидания.

$$\sigma = \sqrt{D(x)}$$

Признаки нормального закона распределения:

1) большинство значений случайной величины группируется около математического ожидания; вероятность появления значения случайной величины, равного математическому ожиданию (среднему) максимальна;

2) вероятность появления наименьших и наибольших значений случайной величины относительно мала.

Плотность вероятности случайной величины, распределенной по нормальному закону

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - M(x))^2}{2\cdot\sigma^2}},$$

где $f(x)$ – функция нормального распределения (плотность распределения вероятности); *M(x)* – математическое ожидание; σ – среднее квадратическое отклонение.

Этапы выборочного исследования:

1) формирование выборки;

2) проведение измерений:

а) составление статистических рядов (простого ряда, ряда распределения, гистограммы и т.п.);

б) подсчет числовых характеристик выборки:

n – объем выборки;

$\overline{X}$ – выборочное среднее (математическое ожидание выборки);

σ – среднее квадратическое отклонение выборки;

3) оценка генерального среднего на основании выборочных показателей:

а) точечная оценка $M \approx \overline{X}$;

б) составление доверительного интервала.

Доверительный интервал – интервал, в который с заранее заданной исследователем доверительной вероятностью попадает генеральное среднее

$$\overline{X} - t \cdot \sigma_{\bar{x}} < M < \overline{X} + t \cdot \sigma_{\bar{x}}$$

M – генеральное среднее (математическое ожидание генеральной совокупности;

t – коэффициент Стьюдента; находится по таблице на основании доверительной вероятности и объема выборки;

$\sigma_{\bar{x}}$ – ***ошибка репрезентативности*** (представительности);

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}(1 - \frac{n}{N})$, где N – объем генеральной совокупности;

если $n << N$ $\qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

если $n = N$ $\qquad \sigma_{\bar{x}} = 0$

I. Работа с одной выборкой

Постановка задачи: пусть получена выборка $x_1, x_2, x_3 \ldots x_n$ объемом n из нормальной генеральной совокупности случайной величины X. Выборочное среднее равно $\overline{X}$; генеральное среднее – M.

Требуется проверить нулевую гипотезу: различие между средними носит случайный характер.

Формула для вычисления t фактического

$$t_{ф} = \frac{\left|\overline{X} - M\right|}{\sigma_{\bar{x}}},$$

где $\sigma_{\bar{x}}$ – ошибка репрезентативности; $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

II. Работа с двумя выборками

Постановка задачи: получены 2 независимые выборки (это так называемые *экспериментальная* выборка (выборка, в которой применяется новая методика) и *контрольная* выборка (выборка, в которой применяется традиционная методика) из нормально распределенной генеральной совокупности случайных величин. Объемы выборок n_1 и n_2; выборочные средние $\overline{X}_1$, $\overline{X}_2$.

Требуется проверить нулевую гипотезу: различие между средними носит случайный характер.

Формула для вычисления t фактического

для большой выборки $n > 30$ $$t_ф = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\sigma_{\bar{x}1}^2 + \sigma_{\bar{x}2}^2}},$$ где $\sigma_{\bar{x}1} = \frac{\sigma_1}{\sqrt{n_1}}$; $\sigma_{\bar{x}2} = \frac{\sigma_2}{\sqrt{n_2}}$;

Алгоритм действий:

1) перед началом исследования с задаваемой вероятностью выдвигаем ***нулевую гипотезу***, предположение о том, что различие между средними носит случайный характер.

2) по таблице коэффициентов Стьюдента на основании вероятности и объема выборки ($n \to \infty$) находится $t_{ст}$ (t стандартное);

3) по определенным формулам вычисляется $t_ф$ (t фактическое);

4) если $t_ф \leq t_{ст}$, то ***нулевая гипотеза подтверждается***, различие носит случайный характер, средние принадлежат одной совокупности; если $t_ф > t_{ст}$, ***нулевая гипотеза отвергается***, принимается альтернативная гипотеза; различие носит существенный характер, средние принадлежат разным совокупностям.