9.2. Критерии проверки гипотез о законах распределения случайной величины.

9.2.1. Критерий А.Н. Колмогорова

Критерий А.Н. Колмогорова применяется для проверки простой гипотезы H_0 о том, что независимые одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$ имеют заданную непрерывную функцию распределения $F(x)$.

Требуется принять или отклонить эту гипотезу по реализации случайной выборки независимых измерений. Для решения этой задачи введем статистику *T* (X^n) критерия проверки гипотезы *H*. Реализация *t* статистики *T* соответствующая выборке x^n определяется по формуле

$$t = \sup_{(x)} \left| F^*(x) - F(x) \right|$$

Доказано, что (H – истинна)$\Rightarrow$(T=D). Здесь D – случайная величина, распределенная по известному закону Колмогорова. Для этой величины можно найти t_α из условия:

$$P(D \geq t_\alpha) = \alpha, \qquad (*)$$

где α - вероятность практически невозможного события, и следовательно, событие ($D \geq t_\alpha$) - практически невозможное.

С точностью до принципа практической уверенности имеем:

$$(H \text{ – истинна}) \Rightarrow (t < t_\alpha);$$

$$(t \geq t_\alpha) \Rightarrow (H \text{ – ложна}).$$

Из этих соотношений следует, что неравенство ($t<t_\alpha$) необходимо для принятия, а неравенство ($t \geq t_\alpha$) достаточно для отклонения гипотезы H (с точностью до принципа практической уверенности).

Руководствуясь этими соображениями, принимают следующее правило решения поставленной задачи:

$$(t < t_\alpha) \Rightarrow (H \text{ – принять});$$

$$(t \geq t_\alpha) \Rightarrow (H \text{ – отклонить}).$$

Это правило называют критерием согласия Колмогорова проверки гипотезы о непрерывной функции распределения случайной величины.

Алгоритм:

1) Провести независимые n-кратные измерения СВ X с непрерывной функцией распределения и получить выборку x^n;

2) Исключить из выборки грубые ошибки;

3) Построить реализацию $F^*(x)$ статистической ФР;

4) Выдвинуть гипотезу $F(x)$ о ФР СВ X;

5) Вычислить параметр t.

6) Задать вероятность α практически невозможного события и из таблицы распределения Колмогорова найти параметр t_α как решение уравнения (*).

7) Принять или отклонить гипотезу

$H=(X \in F(x))$ по решающему правилу.

Доказано, что критерий А.Н. Колмогорова состоятельный и в общем случае смещенный.

Он более чувствителен к различию гипотез, поэтому при прочих равных условиях может применяться для меньших объемов выборки. Поскольку результат проверки признака критерия t зависит от наибольших различий $F(x)$ и $F^*(x)$, то нет необходимости построения $F(x)$ и $F^*(x)$ на всем диапазоне изменения x; достаточно ограничиться областями наибольших различий $F(x)$ и $F^*(x)$.

Недостатком критерия является то, что точность его выводов нарушается, если в формулировании гипотезы о $F(x)$ используются характеристики эмпирических распределений, т.к. в этом случае статистика T зависит от $F(x)$; неудобство доставляет также значительная трудоемкость построения статистики Колмогорова А.Н.

9.2.2. Критерий Пирсона

Критерий Пирсона (критерий χ^2) используется для проверки гипотезы о различных законах распределения с применением статистики:

$$\mathrm{T}(\mathrm{X}^{\mathrm{n}}) = \sum_{j=1}^{q} (N_j - np_j)^2 / np_j = \sum_{j=1}^{q} N_j^2 / np_j - n$$

Здесь N_j – число X_i в разряде статистического ряда, q – число разрядов.

Решающее правило состоит в следующем: если p_j удовлетворяет неравенству

$$\sum_{j=1}^{q} \frac{n_j^2}{np_j} - n \geq t_\alpha$$

то гипотеза H отвергается, в противном случае H принимается.

Алгоритм:

1) по выборке x^n , освобожденной от ошибок, строим статистический ряд, предварительно задав число разрядов q и установив границы разрядов;
2) задав гипотезу о функции распределения или плотности распределения, определяем гипотетические вероятности разрядов

$$p_j = F(x^j - x^{j-1}) = \int_{j-1}^{j} f(x)dx$$

3) вычисляем реализацию $t=T(x^n)$ статистики $T(X^n)$;

4) задавая уровень значимости α, при помощи табл. χ^2 – распределения находим t_α;

5) применяем решающее правило, если $(t \geq t_\alpha)$, то H отклоняем, в противном случае H принимаем.

Достоинства:

- относительная простота;
- возможность применения для векторной X;
- состоятельность;

- возможность применения оценок параметров при формулировании гипотезы *H* без потерь точности выводов;
- несмещенность при p_j=const;
- пониженная требовательность к точности x_i.

Недостатки:

- потери информации за счет предварительного группирования данных по разрядам;
- неопределенность в выборе q и границ разрядов;

- неучет знака разности $N_j - np_j$.

9.2.2.1. Проверка гипотезы о нормальном распределении

Пусть получена выборка достаточно большого объема n с большим количеством различных значений вариант. Для удобства ее обработки, разделим интервал от наименьшего до наибольшего из значений вариант на q равных частей.

Будем считать, что значения вариант, попавших в каждый интервал, приближенно равны числу, задающему середину интервала. Подсчитав число вариант, попавших в каждый интервал, составим так называемую сгруппированную выборку:

варианты…………$x_1 \quad x_2 \quad \ldots \quad x_s$

частоты…………..$n_1 \quad n_2 \quad \ldots \quad n_s$,

где x_i – значения середин интервалов, а n_i – число вариант, попавших в i-й интервал (эмпирические частоты).

По полученным данным можно вычислить выборочное среднее и выборочное среднее квадратическое отклонение σ_B. Проверим предположение, что генеральная совокупность распределена по нормальному закону с параметрами $M(X) =$, $D(X) =$. Тогда можно найти количество чисел из выборки объема n, которое должно оказаться в каждом интервале при этом предположении (то есть теоретические частоты).

Для этого по таблице значений функции Лапласа найдем вероятность попадания в *i*-й интервал:

$$p_i = \Phi\left(\frac{b_i - \bar{x}_B}{\sigma_B}\right) - \Phi\left(\frac{a_i - \bar{x}_B}{\sigma_B}\right),$$

где a_i и b_i - границы *i*-го интервала. Умножив полученные вероятности на объем выборки *n*, найдем теоретические частоты: $n_i = n \cdot p_i$.

Наша цель – сравнить эмпирические и теоретические частоты, которые, конечно, отличаются друг от друга, и выяснить, являются ли эти различия несущественными, не опровергающими гипотезу о нормальном распределении исследуемой случайной величины, или они настолько велики, что противоречат этой гипотезе.

Для этого используется критерий в виде случайной величины

$$\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}$$

Вне зависимости от реального закона распределения генеральной совокупности закон распределения случайной величины при стремится к закону распределения с числом степеней свободы $k = q - 1 - r$, где r – число параметров предполагаемого распределения, оцененных по данным выборки. Нормальное распределение характеризуется двумя параметрами, поэтому $k = q - 3$.

Для выбранного критерия строится правосторонняя критическая область, определяемая условием

$$p(\chi^2 > \chi^2_{kp}(\alpha, k)) = \alpha,$$

где α – уровень значимости. Следовательно, критическая область задается неравенством

$$\chi^2 > \chi^2_{kp}(\alpha, k),$$

а область принятия гипотезы – $\chi^2 < \chi^2_{kp}(\alpha, k)$.

Итак, для проверки нулевой гипотезы H_0: генеральная совокупность распределена нормально – нужно вычислить по выборке наблюдаемое значение критерия:

$$\chi^2_{\text{набл}} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'} \qquad (*)$$

а по таблице критических точек распределения χ^2 найти критическую точку $\chi^2_{\text{кр}}(\alpha, k)$, используя известные значения α и $k = q - 3$. Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ – нулевую гипотезу принимают,

при $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$ ее отвергают.