Проверка статистических гипотез. Элементы теории корреляции

Проверка статистических гипотез

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений.

Например, статистическими являются гипотезы:

1) генеральная совокупность распределена по закону Пуассона;

2) дисперсии двух нормальных совокупностей равны между собой.

В первой гипотезе сделано предположение о виде неизвестного распределения, во второй – о параметрах двух известных распределений.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза.

Нулевой (основной) называют выдвинутую гипотезу H_0.

Конкурирующей (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

Например, если нулевая гипотеза состоит в предположении, что математическое ожидание a нормального распределения равно 10, то конкурирующая гипотеза, в частности, может состоять в предположении, что $a \neq 10$ или $a > 10$. Коротко это записывают так: $H_0 : a = 10$; H_1: $a \neq 10$ или H_1: $a > 10$.

Различают гипотезы, которые содержат только одно и более одного предположений.

Простой называют гипотезу, содержащую только одно предположение. Например, если λ – параметр показательного распределения, то гипотеза $H_0 : \lambda = 5$ – простая. Гипотеза H_0: математическое ожидание нормального распределения равно 3 – простая.

Сложной называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез. Например, сложная гипотеза $H_0 : \lambda > 5$ состоит из бесчисленного множества простых вида $H_1 : \lambda = b_i$, где b_i любое число, большее 5. Гипотеза H_0: математическое ожидание нормального распределения не равно 3 – сложная.

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Поскольку проверку производят статистическими методами, ее называют ***статистической***. В итоге ста-

тистической проверки гипотезы в двух случаях может быть принято неправильное решение, т.е. могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза.

Ошибка второго рода состоит в том, что будет принята неправильная гипотеза.

Подчеркнем, что последствия ошибок могут оказаться весьма различными. Например, если отвергнуто правильное решение "продолжать строительство жилого дома", то эта ошибка первого рода повлечет материальный ущерб; если же принято неправильное решение "продолжать строительство", несмотря на опасность обвала стройки, то эта ошибка второго рода может повлечь гибель людей. Можно привести примеры, когда ошибка первого рода влечет более тяжелые последствия, чем ошибка второго рода.

Замечание 1. Правильное решение может быть принято также в двух случаях:

1) гипотеза принимается, причем в действительности она правильная;

2) гипотеза отвергается, причем в действительности она неверна.

Замечание 2. Вероятность совершить ошибку первого рода принято обозначать через α ; ее называют ***уровнем значимости***. Наиболее часто уровень значимости принимают равным 0,05 или 0,01. Если, например, принят уровень значимости, равный 0,05, то это означает, что в пяти случаях из ста имеется риск допустить ошибку первого рода (отвергнуть правильную гипотезу).

Основной прием проверки статистических гипотез заключается в том, что по имеющейся выборке вычисляется значение некоторой случайной величины, имеющей известный закон распределения.

Статистическим критерием называется случайная величина K с известным законом распределения, служащая для проверки нулевой гипотезы.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (областью допустимых значений) называют совокупность значения критерия, при которых гипотезу принимают.

Критическими точками (границами) $\kappa_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Итак, процесс проверки гипотезы состоит из следующих этапов:

1) выбирается статистический критерий K ;

2) вычисляется его наблюдаемое значение $\kappa_{набл}$ по имеющейся выборке;

3) поскольку закон распределения K известен, определяется (по известному уровню значимости α) ***критическое значение*** $\kappa_{кр}$, разделяющее критическую область и область принятия гипотезы (например, если $p\left(K > \kappa_{кр}\right) = \alpha$, то справа от $\kappa_{кр}$ располагается критическая область, а слева – область принятия гипотезы);

4) если вычисленное значение $\kappa_{набл}$ попадает в область принятия гипотезы, то нулевая гипотеза принимается, если в критическую область – нулевая гипотеза отвергается.

Различают одностороннюю (правостороннюю или левостороннюю) и двусторонние критические области.

Правосторонней называют критическую область, определяемую неравенством $K > \kappa_{кр}$, где $\kappa_{кр}$ – положительное число (рис. 45.1, а).

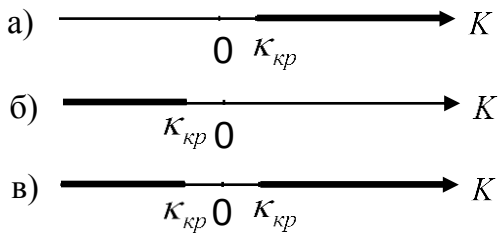

Рис. 45.1.

Левосторонней называют критическую область, определяемую неравенством $K < \kappa_{кр}$, где $\kappa_{кр}$ – отрицательное число (рис. 45.1, б).

Односторонней называют правостороннюю или левостороннюю критическую область.

Двусторонней называют критическую область, определяемую неравенствами $K < \kappa_1$, $K > \kappa_2$, где $\kappa_2 > \kappa_1$.

В частности, если критические точки симметричны относительно нуля, двусторонняя критическая область определяется неравенствами (в предположении, что $\kappa_{кр} > 0$): $K < -\kappa_{кр}$, $K > \kappa_{кр}$, или равносильным неравенством $\left|K\right| > \kappa_{кр}$ (рис. 45.1, в).

Мощностью критерия называют вероятность попадания критерия в критическую область при условии, что верна конкурирующая гипотеза.

Если обозначить вероятность ошибки второго рода (принятия неправильной нулевой гипотезы) β, то мощность критерия равна 1 – β. Следовательно, чем больше мощность критерия, тем меньше вероятность совершить ошибку вто-

рого рода. Поэтому после выбора уровня значимости следует строить критическую область так, чтобы мощность критерия была максимальной.

Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых известны (независимые выборки)

Пусть выдвинуты следующие гипотезы о параметрах распределения: $H_0: M(X) = M(Y)$; $H_1: M(X) > M(Y)$, где X, Y – нормальные генеральные совокупности, выборки из них объемами n и m $(n > 30, m > 30)$ имеют выборочные средние $\bar{x}_B$, $\bar{y}_B$, генеральные дисперсии $D(X)$, $D(Y)$. В качестве критерия проверки нулевой гипотезы примем случайную величину $Z = \dfrac{\bar{x}_B - \bar{y}_B}{\sqrt{D(X)/n + D(Y)/m}}$. (45.1)

При данной конкурирующей гипотезе H_1 строят правостороннюю критическую область $P(Z > z_{кр}) = \alpha$, α – уровень значимости. Находят критическую точку по таблице функции Лапласа из равенства $\Phi(z_{кр}) = 0{,}5(1 - 2\alpha)$. При вычисленном по выборкам $Z_{набл} > z_{кр}$ гипотезу H_0 отвергают. Если же $Z_{набл} < z_{кр}$, то нет оснований отвергать нулевую гипотезу.

Если конкурирующая гипотеза $H_1: M(X) < M(Y)$, то строят левостороннюю критическую область $P(Z < -z_{кр}) = \alpha$, находят $z_{кр}$ из равенства $\Phi(z_{кр}) = 0{,}5(1 - 2\alpha)$, используя таблицу функций Лапласа. Если $Z_{набл} < -z_{кр}$, то нулевую гипотезу отвергают.

Если конкурирующая гипотеза $H_1: M(X) \neq M(Y)$, то строят двухстороннюю критическую область. При $Z_{набл} \in (-z_{кр}; z_{кр})$ гипотеза H_0 принимается. Критическую точку находят из равенства $\Phi(z_{кр}) = 0{,}5(1 - \alpha)$, используя таблицу функции Лапласа.

Пример. В результате исследования накопления содержания натрия в водоёме по двум независимым выборкам, объёмы которых соответственно равны $n = 9$ и $m = 10$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}_B = 25$ (мг/л) (выборка взята в области питания) и $\bar{y}_B = 38$ (мг/л) (выборка взята ниже по течению). Генеральные дисперсии известны: $D(X) = 3{,}6$ (мг/л)^2; $D(Y) = 6$ (мг/л)^2. При уровне значи-

мости $\alpha = 0{,}01$ проверить нулевую гипотезу $H_0\colon M(X) = M(Y)$, при конкурирующей гипотезе $H_1\colon M(X) < M(Y)$.

◁ Подставив данные задачи в формулу (45.1), получим

$$Z_{набл} = \frac{25-38}{\sqrt{\frac{3{,}6}{9}+\frac{6}{10}}} = \frac{-13}{\sqrt{0{,}4+0{,}6}} = -13.$$

По условию, конкурирующая гипотеза имеет вид $M(X) < M(Y)$, поэтому критическая область – левосторонняя.

Найдем "вспомогательную" точку $Z_{кр}$ из равенства $\Phi(Z_{кр}) = 0{,}5(1-2\alpha) = 0{,}5(1-2\cdot 0{,}01) = 0{,}49$.

По таблице функции Лапласа находим $Z_{кр} = 2{,}33$. Следовательно, $Z'_{кр} = -Z_{кр} = -2{,}33$.

Так как $Z_{набл} = -13 < Z_{кр} = -2{,}33$, то нулевую гипотезу отвергаем. Другими словами, выборочная средняя $\bar{x}_B$ значимо меньше выборочной средней $\bar{y}_B$, т.е. в водах происходит накопление содержания натрия. ▷

Элементы теории корреляции

Рассмотрим достаточно большую выборку двумерной случайной величины (*X,Y*), данные которой сгруппированы в виде корреляционной таблицы:

Y	*X*			
	x_1	x_2	…	x_k
y_1	n_{11}	n_{21}	…	n_{k1}
y_2	n_{12}	n_{22}	…	n_{k2}
…	…	…	…	…
y_m	n_{1m}	n_{2m}	…	n_{km}

Здесь n_{ij} – число появлений в выборке пары чисел (x_i, y_j).

Для оценки степени взаимного влияния *X* и *Y* используют коэффициент корреляции.

Выборочным коэффициентом корреляции r_B называется величина $r_B = \dfrac{\overline{(xy)}_B - \overline{x_B}\cdot\overline{y_B}}{\sigma_x \cdot \sigma_y}$, где $\overline{(xy)}_B = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} x_i y_j}{n}$, $\bar{x}_B$, $\bar{y}_B$ – выборочные сред-

ние признаков X и Y, σ_x, σ_y – соответствующие выборочные средние квадратические отклонения.

Замечание 1. Выборочный коэффициент корреляции находится в пределах от -1 до 1, т.е. $-1 \le r_B \le 1$

Замечание 2. $|r_B| = 1$ тогда и только тогда, когда между значениями X и Y имеется линейная зависимость. Чем ближе коэффициент корреляции к нулю, тем хуже эта зависимость выражается линейно.

Пусть выборочный коэффициент корреляции r_B оказался не равен нулю. Это ещё не означает, что и коэффициент корреляции генеральной совокупности не равен нулю. Поэтому при заданном уровне значимости α возникает необходимость проверки нулевой гипотезы $H_0 : r_B = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1 : r_B \neq 0$. Таким образом, при принятии нулевой гипотезы X и Y не коррелированы, а при отклонении H_0 наблюдается корреляционная зависимость.

В качестве критерия примем случайную величину $T = \dfrac{r_B \cdot \sqrt{n-2}}{\sqrt{1-r_B^2}}$, которая при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n-2$ степенями свободы. Из вида конкурирующей гипотезы следует, что критическая область двусторонняя с границами $\pm t_{кр}$, где значение $t_{кр}(\alpha, \kappa)$ находится из таблиц для двусторонней критической области.

Вычислив наблюдаемое значение критерия $T_{набл.} = \dfrac{r_B \cdot \sqrt{n-2}}{\sqrt{1-r_B^2}}$ и сравнив его с $t_{кр}$, делаем вывод:

если $|T_{набл.}| < t_{кр}$ – нулевая гипотеза принимается (корреляции нет);

если $|T_{набл.}| > t_{кр}$ – нулевая гипотеза отвергается (корреляция есть).

Если установлено наличие корреляционной зависимости, то иногда возникает необходимость в нахождении уравнения прямой линии среднеквадратической регрессии.

Уравнение вида $\overline{y_x} = r_B \cdot \dfrac{\sigma_y}{\sigma_x}\left(x - \overline{x_B}\right) + \overline{y_B}$ называется ***уравнением прямой регрессии Y на X.***

Аналогично $\overline{x_y} = r_B \cdot \dfrac{\sigma_x}{\sigma_y}\left(y - \overline{y_B}\right) + \overline{x_B}$ – выборочное ***уравнение прямой регрессии X на Y.***

Число $\rho_{yx} = r_B \cdot \frac{\sigma_y}{\sigma_x}$ называют ***линейным коэффициентом регрессии Y на X.*** Аналогично, $\rho_{xy} = r_B \cdot \frac{\sigma_x}{\sigma_y}$ – ***линейный коэффициент регрессии X на Y.***