Сравнение двух случайных выборок (нахождение оценок параметров распределения, проверка статистических гипотез)

Математические выкладки

Точечные оценки параметров распределения

Распределение случайной величины X характеризуется рядом *параметров* (математическое ожидание, дисперсия и т.д.). Эти параметры называют *параметрами генеральной совокупности*. Важной задачей математической статистики является нахождение по случайной выборке приближенных значений каждого из параметров, называемых *точечными оценками параметров*, или просто *оценками*. Таким образом, *оценкой параметра* β называется функция $f(X_1, X_2, \dots, X_n)$ от случайной выборки, значение которой принимается в качестве приближенного для данного параметра и обозначается $\tilde{\beta}$:

$$\beta \approx \tilde{\beta} = f(X_1, X_2, \dots, X_n). \tag{3.1}$$

Пусть задана повторная случайная выборка $X_1, X_2, \dots, X_n$. *За оценку математического ожидания a* принимается среднее арифметическое элементов выборки:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.2}$$

Оценкой дисперсии σ^2 при *неизвестном математическом ожидании* является величина S^2, которую называют *эмпирической дисперсией*:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3.3}$$

Оценкой среднего квадратического отклонения σ при этом является, соответственно, величина

$$S = \sqrt{S^2} \tag{3.4}$$

Для практических расчетов формулу (3.3) целесообразно преобразовать к следующему виду:

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \tag{3.5}$$

Вычисление среднего значения $\overline{X}$ и оценки дисперсии S^2 упрощается, если отсчет значений X_i вести от подходящим образом выбранного начала отсчета C и в подходящем масштабе, т.е. сделать линейную замену (кодирование):

$$X_i = C + hU_i \ (i = 1, 2, \dots, n). \tag{3.6}$$

При такой замене формулы (3.2), (3.3), (3.4) принимают следующий вид:

$$\overline{X} = C + h\overline{U}; \quad \overline{U} = \frac{1}{n}\sum_{i=1}^{n} U_i. \tag{3.7}$$

$$S^2 = \frac{h^2}{n-1}\sum_{i=1}^{n}(U_i - \overline{U})^2 = \frac{h^2}{n-1}\left(\sum_{i=1}^{n} U_i^2 - n\overline{U}^2\right). \tag{3.8}$$

Для контроля правильности вычислений весь расчет следует повторить при другом начале отсчета C: результаты должны совпадать с точностью до величины возможных ошибок округления.

Доверительные интервалы параметров нормального распределения

Доверительным интервалом параметра β называется интервал со случайными границами ($\tilde{\beta} - \varepsilon_1$; $\tilde{\beta} + \varepsilon_2$), который накрывает истинное значение параметра β с заданной вероятностью P, которая называется *доверительной вероятностью*. Величина $\alpha = 1 - P$ называется *уровнем значимости*. При этом обычно требуют, чтобы вероятности выхода за границы доверительного интервала в обе стороны были равны между собой, а именно:

$P(\beta < \tilde{\beta} - \varepsilon_1) = P(\beta > \tilde{\beta} + \varepsilon_2) = (1 - P)/2 = \alpha/2$.

Это дополнительное требование обеспечивает единственность решения задачи.

Пусть задана повторная случайная выборка $X_1, X_2, ..., X_n$ из нормальной генеральной совокупности. Это означает, что результаты эксперимента независимы и подчиняются нормальному закону распределения с одинаковыми параметрами $X_i \sim N(a; \sigma)$.

С вероятностью P математическое ожидание принадлежит интервалу

$$a \in (\overline{X} - \varepsilon,\ \overline{X} + \varepsilon); \tag{3.9}$$

$$\varepsilon = t_{1-\alpha/2}(k)\, S/\sqrt{n} \tag{3.10}$$

где $\overline{X}$ – оценка математического ожидания (3.2); $S = \sqrt{S^2}$ – оценка среднего квадратического отклонения σ (3.4);

$t_{1-\alpha/2}(k)$ – квантиль распределения Стьюдента с k степенями свободы; n – объем выборки; k – число степеней свободы при вычислении оценки S.

Часто доверительный интервал для математического ожидания записывают символически:

$$a = \overline{X} \pm \varepsilon. \tag{3.11}$$

Доверительный интервал для среднего квадратического отклонения σ при доверительной вероятности $P = 1 - \alpha$ имеет следующий вид:

$$S\sqrt{\frac{k}{\chi^2_{1-\alpha/2}(k)}} < \sigma < S\sqrt{\frac{k}{\chi^2_{\alpha/2}(k)}}, \tag{3.12}$$

где S – оценка среднего квадратического отклонения σ (1.6) при неизвестном математическом ожидании; $\chi_P^2(k)$ – квантиль распределения Пирсона с k степенями свободы; k – число степеней свободы оценки S.

Проверка статистических гипотез

Пусть X – наблюдаемая случайная величина. *Статистической гипотезой* H называется предположение относительно параметров или вида распределения случайной величины X. Проверяемая гипотеза называется *нуль-гипотезой* и обозначается H_0. При постановке нуль-гипотезы сразу ставится *альтернативная гипотеза* H_1, т.е. то предположение, которое следует принять, если нуль-гипотеза будет отвергнута.

Правило, позволяющее принять или отвергнуть гипотезу H_0 – некоторая функция результатов эксперимента $Q(X_1, X_2, ..., X_n)$, *распределение которой вполне определено при условии истинности гипотезы* H_0.

Пусть заданы две независимые выборки из двух нормальных генеральных совокупностей. Первая выборка имеет объем n_1, элементы выборки $X_i^{(1)} \sim N(a_1; \sigma_1)$; вторая – объем n_2, элементы выборки $X_i^{(2)} \sim N(a_2; \sigma_2)$. Необходимо проверить *гипотезу о равенстве дисперсий* этих двух генеральных совокупностей, т.е. H_0: $\sigma_1^2 = \sigma_2^2$. Математические ожидания a_1 и a_2 неизвестны.

В этом случае по каждой выборке находят несмещенные оценки дисперсий S_1^2 и S_2^2 с числами степеней свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ соответственно. Гипотезу проверяют по критерию Фишера, функция критерия

$$F = S_1^2/S_2^2 \tag{3.13}$$

имеет F-распределение Фишера с k_1 и k_2 степенями свободы, т.е. $F = F(k_1,k_2)$.

Если альтернативная гипотеза H_1: $\sigma_1^2 \neq \sigma_2^2$, то критерий Фишера рассчитывается как отношение большей по величине оценки дисперсии к меньшей:

$$F = (S_{бол})^2/(S_{мен})^2 > 1. \tag{3.14}$$

Гипотеза принимается при выполнении неравенства

$$F < F_{1-\alpha/2}(k_{S_{бол}},k_{S_{мен}}), \tag{3.15}$$

в противоположном случае гипотеза отвергается. Здесь $k_{S_{бол}}$ – число степеней свободы большей оценки дисперсии; $k_{S_{мен}}$ – число степеней свободы меньшей оценки дисперсии.

Если гипотеза о равенстве дисперсий принимается, то за оценку общей σ может быть взята $S = \sqrt{S_{св}^2}$, полученная по формуле для сводной оценки дисперсии:

$$S_{св}^2 = \frac{k_1 S_1^2 + k_2 S_2^2}{k_1 + k_2}; \quad k_j = n_j - 1; \quad j = 1, 2, \tag{3.16}$$

где S_1^2, S_2^2 – несмещенные оценки дисперсии первой и второй выборок соответственно.

Пусть заданы две независимые выборки из двух нормальных генеральных совокупностей.

Первая выборка имеет объем n_1, элементы выборки $X_i^{(1)} \sim N(a_1; \sigma_1)$; вторая – объем n_2, элементы выборки $X_i^{(2)} \sim N(a_2; \sigma_2)$. Математические ожидания a_1 и a_2 неизвестны.

Проверяем *гипотезу о равенстве математических ожиданий* этих двух генеральных совокупностей, т.е. H_0: $a_1 = a_2$. По каждой выборке находим оценки математических ожиданий $\overline{X}_1$ и $\overline{X}_2$. При этом дисперсии σ_1^2 и σ_2^2 неизвестны, но гипотеза о равенстве дисперсий принимается; S_1^2 и S_2^2 – несмещенные оценки дисперсий первой и второй выборок. Находим сводную оценку дисперсии (3.16).

Гипотеза проверяется по критерию Стьюдента, функция критерия

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{св}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.17}$$

имеет t-распределение Стьюдента с k_{CB} степенями свободы, т.е. $t = t(k_{CB})$; $k_{CB} = k_1 + k_2$ – число степеней свободы при вычислении оценки $S_{СВ} = \sqrt{S_{СВ}^2}$. При альтернативной гипотезе H_1: $a_1 \neq a_2$, гипотеза принимается при выполнении неравенства

$$|t| < t_{1-\alpha/2}(k_{CB}), \tag{3.18}$$

в противоположном случае гипотеза отвергается.

Если гипотеза о равенстве математических ожиданий принимается, то за оценку общего математического ожидания может быть взята сводная оценка математического ожидания, которая определяется как средняя по обеим выборкам:

$$\overline{X}_{СВ} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}. \tag{3.19}$$

Доверительный интервал для математического ожидания при этом можно пересчитать, заменив в формулах (3.10) – (3.11) $\overline{X}$ на $\overline{X}_{CB}$; S на S_{CB}; k на $k_{CB} = k_1 + k_2$, n на $n_1 + n_2$.

Проверка гипотезы о виде распределения генеральной совокупности

Если распределение случайной величины X не известно, можно рассмотреть гипотезу о том, что X имеет функцию распределения $F(x)$. Критерии значимости для проверки таких гипотез называются *критериями согласия*.

Пусть $X_1, X_2, \ldots, X_n$ – выборка наблюдений случайной величины X. Проверяется гипотеза H_0, утверждающая, что X имеет функцию распределения $F(x)$.

Проверку гипотезы H_0 при помощи критерия χ^2 проводят следующим образом. По выборке находят оценки неизвестных параметров предполагаемого закона распределения случайной величины X. Область возможных значений случайной величины X разбивают на l интервалов. Подсчитывают числа n_i попаданий результатов экспериментов в каждый i-й интервал. Используя предполагаемый закон распределения случайной величины X, находят вероятности p_i того, что значение X принадлежит i-му интервалу. Затем сравнивают полученные частоты n_i/n с вероятностями p_i. Критерий согласия Пирсона требует принятия гипотезы о пригодности проверяемого распределения с уровнем значимости α, если значение *взвешенной суммы квадратов отклонений*

$$\chi^2 = n\sum_{i=1}^{l}\frac{(n_i/n-p_i)^2}{p_i} = \sum_{i=1}^{l}\frac{(n_i-np_i)^2}{np_i}\,. \tag{3.20}$$

меньше квантиля распределения χ^2-распределения с $k = l - 1$ степенями свободы, т.е. $\chi^2 < (\chi_{1-\alpha})^2(k)$,

в противоположном случае эта гипотеза отвергается, как противоречащая результатам эксперимента. Если при этом некоторые параметры распределения оценивают по результатам той же выборки, то квантиль χ^2-распределения следует брать для $k = l - 1 - m$ степеней свободы, где m – число оцениваемых параметров.

Построение гистограммы (эмпирической функции распределения)

Для наглядного представления о выборке часто используют график, называемый *гистограммой*. Для построения гистограммы интервал, содержащий все элементы выборки, разбивают на l непересекающихся интервалов (как правило, равной длины). Подсчитывают числа n_i попаданий результатов экспериментов в каждый i-й интервал и строят столбиковую диаграмму, откладывая по оси ординат значения средней плотности $n_i/(nh_i)$, где h_i – длина i-го интервала. Площадь каждого столбика равна n_i/n, что соответствует относительной частоте попадания элементов выборки в i-й интервал. Площадь под всей ступенчатой фигурой равна единице.

При увеличении объема выборки и уменьшении интервалов группировки гистограмма приближается к функции плотности генеральной совокупности. Гистограмма является эмпирической функцией плотности, она дает приближенную функцию плотности генеральной совокупности (ее оценку) по случайной выборке.