KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 5 урок > Непараметрические статистические критерии

> Непараметрические критерии. Критерий знаков

До этого момента мы рассматривали параметрические критерии, которые работают только для выборок с известным распределением. Что же делать, если распределение нашей выборки неизвестно?

Существует несколько методов решения этой задачи:

- Критерий знаков
- Критерий рангов
- Критерий Манна-Уитни-Вилкоксона

Критерий знаков

В случае отсутствия информации о распределении нашей выборки, можно перейти от неё к другой выборке с уже известным распределением.

Предположим, что число m является медианой выборки N, тогда можно создать новую выборку O, сравнив каждое значение из выборки N с m. В случае, если конкретное значение оказалось больше медианы, соответсвующее значение в новой выборке O будет 1, а в случае, если меньше медианы — 0. Если m действительно является медианой, то новая выборка O будет иметь биномиальное распределение.

В случае одной выборки:

- Выборка: $X_1, \ldots, X_N \sim P$ (P не известно)
- Нулевая гипотеза: $median(X) = m$
- Альтернативная гипотеза: $median(X) \neq m$
- Статистика T_N

$T_N = \sum_{x=1}^{N} [X_i > m]$

- Нулевое распределение: $T_N \sim Bin(N, 1/2)$

В случае связанных выборок:

- Выборка: $X_1 1, \ldots, X_1 N, X_2 1, ..., X_2 N$ — связанные выборки
- Нулевая гипотеза: $P(X_1 > X_2) = 1/2$
- Альтернативная гипотеза: $P(X_1 > X_2) \neq 1/2$
- Статистика T_N

$T_N = \sum_{x=1}^{N} [X_{1i} > X_{2i}]$

- Нулевое распределение: $T_N \sim Bin(N, 1/2)$

Однако, превратив выборку в выборку бинарных величин, мы потеряли часть информации и получили более слабый по мощности критерий.

> Критерий рангов. Критерий Манна-Уитни-Уилкоксона

Критерий рангов

Можно реализовать промежуточный вариант — с одной стороны отказаться от абсолютных значений, а с другой сохранить порядок в выборке. Это и будет критерий рангов.

Вариационный ряд — отсортированная по возрастанию выборка. Одинаковые элементы в выборке образуют связки.

- Если X_i не в связке, то $rank(X_i) = r : X_i = X(r)$
- Если X_i в связке от k_1 до k_2, то $rank(X_i) = (k_1 + k_2)/2$

	consume	consume_rank
0	5.0	5.5
1	4.2	2.0
2	5.5	8.0
3	3.9	1.0
4	4.5	4.0
5	6.4	9.5
6	4.4	3.0
7	5.0	5.5
8	6.4	9.5
9	5.3	7.0

Критерий Манна-Уитни-Уилкоксона

Критерий Манна-Уитни-Уилкоксона проверяет, имеют ли две независимые выборки одно распределение или же распределение одной выборки смещено относительно распределения другой выборки.

- $X_{11}, ..., X_{1N}, X_{21}, ..., X_{2N}$ — независимые выборки
- Нулевая гипотеза: $F_{X1}(x) = F_{X2}(x)$
- Альтернативная гипотеза: $F_{X1}(x) = FX2(x + \Delta), \Delta \neq 0$
- Вариационный ряд объединённой выборки X_1 и X_2, статистика R_N:

$R_N = \sum_{i=1}^{N_1} rank(X_{1i})$

- Нулевое распределение: R_N — табличное

Что такое табличное распределения?

Это ситуация, в которой мы просто взяли и посчитали вероятности для всех возможных значений выборки. Нулевая гипотеза предполагает, что распределения не отличаются, значит, в вариационном ряде элементы первой выборки находятся на случайных местах. Можем перебрать все возможные варианты расположения элементов первой выборки в объединённой выборке и посчитать вероятность, то есть сумму рангов
элементов первой выборки в вариационном ряде.

- Таблицы неудобно хранить — нужно посчитать и запомнить для всевозможных N_1, N_2
- Если выборки достаточно большие $N_1, N_2 > 10$, то табличное распределение становится похожим на нормальное:

$$R_N \sim \mathbb{N}(\frac{N_1(N_1+N_2+1)}{2}, \frac{N_1 N_2(N1+N2+1)}{12})$$

> Множественная проверка гипотез. Поправка Борферрони. Метод Холма

Например, мы имеем N моделей, каждая из которых претендует улучшить результаты текущей. Тогда можно оценить шансы ложно принять хотя бы 1 экспериментальную модель:

- Вероятность не отвергнуть 1 модель: $1 - \alpha$
- Вероятность не отвергнуть N моделей: $(1 - \alpha)^N$
- Вероятность отвергнуть хотя бы 1 из N: $1 - (1 - \alpha)^N$
- Для $N = 10, \alpha = 0.05 : P = 0.40126$

Таким образом, при множественной проверке гипотез или моделей вероятность ошибки первого рода растет с увеличением количества гипотез или моделей.

Групповая ошибка первого рода $FWER(V > 0)$, где V— число ложноотвергнутых гипотез. Логично предположить, что её вероятность можно снизить путем изменения уровней значимостей каждой гипотезы α.

Поправка Бонферрони

- Возьмём новые уровни значимости: $\alpha_1 = ... = \alpha_m = \alpha/m$
- Вспомним вероятность отвергнуть хотя бы 1 из N гипотез:

$1 - (1 - \alpha_i)^N = 1 - (1 - \alpha/N)^N$

- Для $N = 10 : 0.04889 \approx 0.05$
- Для $N = 100 : 0.04878 \approx 0.05$

Однако поправка Бонферрони сильно уменьшает мощность тестов, то есть ухудшает возможность детектирования эффекта при его наличии.

Нисходящие методы проверки

В нисходящих методах проверки нужно сначала посчитать уровни значимости для гипотез $H_1, ..., H_m$, а затем отсортировать гипотезы по их уровням значимости:

$p_(1) \leq ...p_(m)$ $\qquad\qquad$ $H_(1), ..., H_(m)$

Проверка будет осуществляться по следующему алгоритму:

1. Если $p_(1) > \alpha_(1)$, то принимаем все нулевые гипотезы $H_(1), ..., H_(m)$, иначе отвергаем $H_(1)$ и проверяем дальше.

2. Если $p_(2) > \alpha_(2)$, то принимаем все нулевые гипотезы $H_(2), ..., H_(m)$, иначе отвергаем $H_(2)$ и проверяем дальше.

3. ...

Метод Холма

Одним из нисходящих методов является метод Холма. Этот метод обеспечивает FWER на уровне α.

$$\alpha_1 = \frac{\alpha}{m}$$

$$\alpha_2 = \frac{\alpha}{m-1}$$

$$\alpha_i = \frac{\alpha}{m - i + 1}$$

$$\alpha_m = \alpha$$

На практике:

- Методы Бонферрони и Холма работают для независимых гипотез, а метрики, для которых мы оцениваем статистические значимости, не являются независимыми!
- Некоторые метрики сонаправлены и коррелируют друг с другом, а мы предполагали, что каждая гипотеза может быть опровергнута независимо.
- Метод Холма мощнее Бонферрони.