Лекция 5

Тестрование гипотез. Непараметрика

Курс: Введение в DS на УБ и МиРА (весна, 2022)
Преподаватель: Владимир Омелюсик

25 апреля 2022 г.

В предыдущих сериях

- Ковариация и корреляция.
- Формулировка основной гипотезы.
- Ошибки I и II рода.

Статистический тест и распределения

- Статистический тест (критерий) – математическое правило, в соответствии с которым отвергается или не отвергается проверяемая гипотеза.
- Бывают параметрическими (проверка параметров распределений) и непараметрическими (независимость).
- Обычно рассчитается какая-то статистика, которая при верной H_0 имеет какое-то распределение. Далее смотрят, насколько вероятно при верной H_0 увидеть такое значение статистики.

Статистическая значимость

- Напоминание: ошибка I рода (False Positive) - ситуация, когда отвергнута верная нулевая гипотеза.
- Хотим, чтобы $\mathbb{P}(\text{ош. I рода}) \leqslant \alpha$, где α выбираем сами. Обычно используются значения 0.01, 0.05, 0.1.

Картинка:

Z-тест для одной выборки

$X_1, \ldots, X_N$ – выборка из N независимых, одинаково распределённых нормальных случайных величин с известной дисперсией.

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

Тогда Z-статистка для гипотезы

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0 \end{cases}$$

рассчитывается по формуле

$$Z_{obs} = \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}} \sim \mathcal{N}(0, 1)$$

t-тест для одной выборки

$X_1, \ldots, X_N$ – выборка из N независимых, одинаково распределённых нормальных случайных величин с неизвестной дисперсией.

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

Тогда t-статистка для гипотезы

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0 \end{cases}$$

рассчитывается по формуле

$$t_{obs} = \frac{\hat{\mu} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{N}}} \sim t_{N-1}$$

NB!: $t_{N-1} \to Z$ при $N \to \infty$.

Пример: проверка гипотезы о среднем

$$X_1, \ldots, X_{300} \sim \mathcal{N}(\mu, \sigma)$$

$$\sum_i X_i = 60, \ \sum_i X_i^2 = 4000$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 0.3, \\ H_1 : \mu \neq 0.3 \end{cases}$$

на уровне значимости 5%.

Для справки: $Z_{0.975} \approx t_{299,0.975} \approx 1.96$

Пример: проверка гипотезы о среднем

Проверка гипотезы о доле

$X_1, \ldots, X_N$ – выборка из N независимых, одинаково распределённых случайных величин Бернулли

$$X_i \sim \mathrm{Bern}(1, p).$$

Тогда Z-статистка для гипотезы

$$\begin{cases} H_0 : p = p_0, \\ H_1 : p \neq p_0 \end{cases}$$

рассчитывается по формуле

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \to \mathcal{N}(0, 1)$$

NB!: Можно проверять только при большом N. t-статистику использовать нельзя.

Пример: проверка гипотезы о доле

$$X_1, \ldots, X_{300} \sim \mathrm{Bern}(1, p)$$

$$\sum_i X_i = 60,$$

Проверьте гипотезу

$$\begin{cases} H_0 : p = 0.3, \\ H_1 : p \neq 0.3 \end{cases}$$

на уровне значимости 5%.

Для справки: $Z_{0.975} \approx 1.96$

Пример: проверка гипотезы о доле

p-value

p-value
Минимальный уровень значимости, при котором нулевая гипотеза не отвергается.

Пусть проверяем гипотезу о равенстве средних или долей против гипотезы о неравенстве. Посчитали Z_{obs} или t_{obs}. Тогда

$$\text{p-value} = 2\,\mathbb{P}(Z \leqslant Z_{obs})$$

Основной результат: p-value $< \alpha \Rightarrow$ **нулевая гипотеза отвергается.**

Замечаение: общая формула теста

Проверка гипотез при помощи p-value

Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 0.3, \\ H_1 : \mu \neq 0.3 \end{cases}$$

на уровне значимости 5% (используйте ± 1.96 как критеческое значение), если

1. Оказалось, что $t_{obs} = -3.14$.
2. Оказалось, что $t_{obs} = 0.02$.

Доверительный интервал
Интервал со случайным границами, такой что

$$\mathbb{P}(T_l(X) < \theta < T_r(X)) \geqslant 1 - \alpha$$

Картинка:

Проверка гипотез при помощи доверительного интервала

Заметим, что H_0 не отвергается на уровне значимости 5% при использовании Z-теста, когда

$$-1.96 \leqslant \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{N}} \leqslant 1.96$$

$$-1.96 \frac{\sigma}{\sqrt{N}} \leqslant \hat{\mu} - \mu_0 \leqslant 1.96 \frac{\sigma}{\sqrt{N}},$$

$$\hat{\mu} - 1.96 \frac{\sigma}{\sqrt{N}} \leqslant \mu_0 \leqslant \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{N}}$$

Основной результат
Если наблюдаемое значение статистики попадает в $(1 - \alpha)$-процентный доверительный интервал, то нулевая гипотеза не отвергается на уровне значимости α.

Две выборки

1. Выборки независимы, равенство средних (долей):
    1.1 Дисперсии известны, либо обе выборки большие.
    1.2 Дисперсии неизвестны, но предполагаем, что равны.
    1.3 Диспресии неизвестны и не предполагаем, что равны (сложно).
2. Выборки зависимы: связанные пары.

1.1

$X_1, \ldots, X_{N1}$ и Y_1, $\ldots$, Y_{N2} – две независимые выборки. Предположим, что дисперсии этих выборок известны или $N1$ и $N2$ велики. Тогда для проверки гипотезы о равенстве средних (долей) можно использовать статистику

$$Z_{obs} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{N1} + \frac{\sigma_Y^2}{N2}}} \sim \mathcal{N}(0, 1).$$

$X_1, \ldots, X_{N1}$ и Y_1, ..., Y_{N2} – две независимые выборки. Предположим, что дисперсии этих выборок неизвестны и равны. Тогда для проверки гипотезы о равенстве средних можно использовать статистику

$$t = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\hat{\sigma}\sqrt{\frac{1}{N1} + \frac{1}{N2}}} \sim t_{N1+N2-2},$$

$$\hat{\sigma} = \sqrt{\frac{(N1-1)\hat{\sigma}_X^2 + (N2-1)\hat{\sigma}_Y^2}{N1+N2-2}}$$

Тесты для парных выборок

Пусть есть N наблюдений над одним объектом ДО и ПОСЛЕ проведения эксперимента: $X_1, \ldots, X_N$ и $Y_1, \ldots, Y_N$. Составим разности $d_i = y_i - x_i$ и проверим гипотезу о равенстве средних ДО и ПОСЛЕ:

$$\begin{cases} H_0 : \mu_d = 0, \\ H_1 : \mu_d \neq 0 \end{cases}$$

Тогда гипотезу о разности средних ДО и ПОСЛЕ можно проверить при помощи статистики

$$t = \frac{\bar{d} - \mu_d}{\frac{\sigma_d}{\sqrt{N}}} \sim t_{N-1},$$

где

$$\sigma_d = \sqrt{\frac{1}{N-1}\left(\sum_i d_i^2 - \frac{(\sum_i d_i)^2}{N}\right)}.$$

Непараметрика: задачи

Непараметрическая статистика
Охватывает методы, которые не полагаются на данные, относящиеся к какому-либо конкретному распределению. При этом параметры (средние, медианы и проч.) присутствуют, но не фиксированы заранее.

Примеры задач:

- Проверка независимости двух выборок (здесь же аналоги корреляций для категориальных переменных).
- Проверка того, пришли ли выборки из одного семейства распределений.
- Гистограмма и ядерная оценка плотности.

Таблица сопряжённости

Сразу пример:

Критерий согласия для проверки независимости

$$\chi^2_{obs} = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \overset{H_0}{\sim} \chi^2_{(R-1)(C-1)},$$

где

- R и C – число строк и столбцов в таблице сопряжённости.
- $O_{i,j}$ – количество наблюдений в клетке (i,j) таблицы сопряжённости.
- N – число наблюдений в выборке.
- $E_{i,j} = Np_{i\cdot}p_{\cdot j}$.
- $p_{i\cdot} = \sum_{j=1}^{C} \frac{O_{i,j}}{N}$.
- $p_{\cdot j} = \sum_{i=1}^{R} \frac{O_{i,j}}{N}$.

Пример: садоводство

На участке 1 собрали 25 яблок низкого качества, 50 яблок среднего качества и 25 яблок высокого качества, а на участке 2 собрали 52 яблока низкого качества, 41 яблоко среднего качества и 7 яблок высокого качества. Существует ли зависимость между типом участка и качеством урожая? Используйте уровень значимости 5%. Для справки: $\chi^2_{2,0.95} = 5.99$.

Пример: садоводство