Теория вероятностей и математическая статистика

Тестирование статистических гипотез I.

Глеб Карпов

ВШБ Бизнес-информатика

Основы тестирования гипотез

Пусть $X_1, \dots, X_n$ — случайная выборка независимых случайных величин X_i с функцией плотности $f_X(x|\theta)$, где θ — неизвестный параметр распределения X_i.

Имя параметра θ — это просто общее обозначение для многих возможных вариантов, мы будем работать с параметрами μ и p, как и раньше.

Задача тестирования гипотез задаётся разбиением пространства параметров Θ на два непересекающихся подмножества: $\Theta = \Theta_0 \cup \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$

Две гипотезы формулируются следующим образом:

$$H_0 : \theta \in \Theta_0 \qquad H_1 : \theta \in \Theta_1$$

где:

- H_0 — нулевая гипотеза
- H_1 — альтернативная гипотеза

Процедура тестирования гипотез

Процедура тестирования гипотез определяет:

- При каких значениях выборки не отклонять H_0
- При каких значениях выборки отклонять H_0 в пользу H_1

Возможные исходы

Четыре возможных исхода в тестировании гипотез:

Решение	H_0 верна	H_0 неверна
Принять H_0	правильно	Ошибка II рода
Отклонить H_0	Ошибка I рода	правильно

Вероятность ошибки I рода = уровень значимости (α).

Обычно $\alpha = 0.05$

Философские замечания

- Никогда не утверждать категорично, что одна из двух гипотез верна
- "Принять H_0" на самом деле означает "Пока что не отклонять H_0"
- Одиночный эксперимент не может выступить абсолютным подтверждением гипотезы

Гипотезы о математическом ожидании при известной дисперсии

Предположения:

1. Случайная выборка $\mathcal{X} = \{X_1, \dots, X_n\}$ из неизвестного распределения с $\mu \equiv E[X_i]$, и $\sigma^2 \equiv \mathrm{Var}\,[X_i]$,
2. Истинную дисперсию считаем известной,
3. Либо: большая выборка ($n > 30$), либо исследуемый случайный процесс имеет нормальное распределение

Гипотезы о математическом ожидании при известной дисперсии

Пример 1: левосторонний тест

Исследовательская команда изучает популяции жирафов. Исследователи верят, что средняя длина шеи жирафов огромная — около 20 метров. Однако в выборке было получено выборочное среднее 2.8 метров. Команда хочет проверить, действительно ли средняя длина шеи меньше того значения, в которое они верят.

- Случайная величина для анализа: X — длина шеи. Дисперсию X предположим известной и правдивой.
- Предполагается, что длины шеи следуют нормальному распределению.

Данные:

- Выборка: $n = 20$ жирафов
- Выборочное среднее: $\bar{x} = 2.8$ метров
- Известная дисперсия X: $\text{Var}[X] = 0.16$.

Гипотезы:

- $H_0 : \mu = 20$
- $H_1 : \mu < 20$

Гипотезы о математическом ожидании при известной дисперсии

Левосторонний тест

Для тестирования $H_0 : \mu = \mu_0$ против $H_1 : \mu < \mu_0$:

- *Распределение при нулевой гипотезе*, когда мы безраздельно верим в то, что она верна:

$$\bar{X} \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$$

- Вычисление критической области:

$$P_{H_0}\left(\bar{X} < K\right) = \alpha$$

- Преобразование:

$$P_{H_0}(\bar{X} < K) = P_{H_0}\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < \frac{K - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = P(Z < z_\alpha)$$

- Граница критической области:

$$K = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$$

- Правило принятия решения: Отклонить H_0, если $\bar{x} < K$

Гипотезы о математическом ожидании при известной дисперсии

Пример 1: решение

- **Гипотезы**: $H_0 : \mu = 20$ против $H_1 : \mu < 20$ (левосторонний тест)
- **Данные**: $\mu_0 = 20$, $\sigma = 0.4$, $n = 20$, $\bar{x} = 2.8$, $\alpha = 0.05$
- **Критическая область**: $z_{0.05} = 1.645$

$$K = 20 - 1.645 \cdot \frac{0.4}{\sqrt{20}} \approx 19.853$$

- **Решение**: $\bar{x} = 2.8 < K = 19.853$, поэтому отклоняем H_0.
- **Вывод**: Имеются достаточно статистически значимые основания для утверждения, что средняя длина шеи жирафов меньше 20 метров. Консервативная гипотеза отклоняется в пользу альтернативной.

Гипотезы о математическом ожидании при известной дисперсии

Пример 2: правосторонний тест

Исследовательская команда изучает популяции жирафов. Исследователи верят, что средняя длина шеи жирафов составляет 0.5 метра. Однако при наблюдении выборки было получено выборочное среднее 2.8 метров. Команда хочет проверить, действительно ли средняя длина шеи больше ожидаемых 0.5 м.

- Случайная величина для анализа: X — длина шеи. Дисперсия X известна из предыдущих исследований.
- Предполагается, что длины шеи следуют нормальному распределению.

Данные:

- Выборка: $n = 20$ жирафов
- Выборочное среднее: $\bar{x} = 2.8$ метров
- Известная дисперсия X: $\text{Var}[X] = 0.01$.

Гипотезы:

- $H_0 : \mu = 0.5$
- $H_1 : \mu > 0.5$

Гипотезы о математическом ожидании при известной дисперсии

Правосторонний тест

Для тестирования $H_0 : \mu = \mu_0$ против $H_1 : \mu > \mu_0$:

- *Распределение при нулевой гипотезе*, когда мы безраздельно верим в то, что она верна:

$$\bar{X} \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$$

- Общая идея: отклонять H_0, если свидетельство $\bar{x}$ достаточно велико, *т.е.* $\bar{x} > K$. Как выбрать K?
- Установление критической (решающей) границы:

$$P_{H_0}\left(\bar{X} > K\right) = \alpha$$

- Преобразование:

$$P_{H_0}(\bar{X} > K) = P_{H_0}\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{K - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = P(Z > z_\alpha)$$

- В итоге, критическая (отклоняющая) граница:

$$K = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

- Правило принятия решения: Отклонить H_0, если $\bar{x} > K$

Гипотезы о математическом ожидании при известной дисперсии

Пример 2: решение

- **Гипотезы**: $H_0 : \mu = 0.5$ против $H_1 : \mu > 0.5$ (правосторонний тест)
- **Данные**: $\mu_0 = 0.5$, $\sigma = 0.1$, $n = 20$, $\bar{x} = 2.8$, $\alpha = 0.05$
- **Критическая область**: $z_{0.05} = 1.645$

$$K = 0.5 + 1.645 \cdot \frac{0.1}{\sqrt{20}} \approx 0.537$$

- **Решение**: $\bar{x} = 2.8 > K = 0.647$, поэтому отклоняем H_0.
- **Вывод**: Имеются достаточно статистически значимые основания для утверждения, что средняя длина шеи жирафов больше 0.5 метра. Консервативная гипотеза отклоняется в пользу альтернативной.

Гипотезы о математическом ожидании при известной дисперсии

Пример 3: двусторонний тест

Интернет-магазин недавно переделал свой сайт для улучшения пользовательского опыта. До переделки среднее время, которое клиенты проводили на сайте, составляло 8 минут. Аналитическая команда компании знает из исторических данных, что стандартное отклонение длительности сессии составляет 2.5 минуты. Команда хочет определить, изменила ли переделка среднюю длительность сессии.
После переделки была проанализирована случайная выборка из 50 сессий клиентов, и средняя длительность сессии оказалась равной 7.2 минуты. Компании нужно определить, является ли это различие статистически значимым, так как любое изменение длительности сессии может повлиять на вовлечённость клиентов и продажи.

- Случайная величина для анализа: X — длительность сессии. Дисперсия X известна из исторических данных.
- Предполагается, что длительности сессий следуют нормальному распределению.

Данные:

- Выборка: $n = 50$ сессий
- Выборочное среднее: $\bar{x} = 7.2$ минут
- Известное стандартное отклонение: $\sigma = 2.5$ минут

Гипотезы:

- $H_0 : \mu = 8$ (переделка не изменила среднюю длительность сессии)
- $H_1 : \mu \neq 8$ (переделка изменила среднюю длительность сессии)

Гипотезы о математическом ожидании при известной дисперсии

Двусторонний тест

Для тестирования $H_0 : \mu = \mu_0$ против $H_1 : \mu \neq \mu_0$:

- *Распределение при нулевой гипотезе*, когда мы безраздельно верим в то, что она верна:

$$\bar{X} \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$$

- Вычисление критической (отклоняющей) области:

$$P_{H_0}\left(\bar{X} > K_R\right) = \frac{\alpha}{2}, \quad P_{H_0}\left(\bar{X} < K_L\right) = \frac{\alpha}{2} \quad \text{или } P_{H_0}\left(K_L < \bar{X} < K_R\right) = 1 - \alpha$$

- Преобразование:

$$P_{H_0}\left(K_L < \bar{X} < K_R\right) = P_{H_0}\left(\frac{K_L - \mu_0}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < \frac{K_R - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = P(-z_{\alpha/2} < Z < z_{\alpha/2})$$

- Границы критической области:

$$z_{\alpha/2} = \frac{K_R - \mu_0}{\frac{\sigma}{\sqrt{n}}} \to K_R = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad -z_{\alpha/2} = \frac{K_L - \mu_0}{\frac{\sigma}{\sqrt{n}}} \to K_L = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

- Правило принятия решения: Отклонить H_0, если $\bar{x} < K_L$ или если $\bar{x} > K_R$.

Гипотезы о математическом ожидании при известной дисперсии

Пример 3: решение

- **Гипотезы**: $H_0 : \mu = 8$ против $H_1 : \mu \neq 8$ (двусторонний тест)
- **Данные**: $\mu_0 = 8$, $\sigma = 2.5$, $n = 50$, $\bar{x} = 7.2$, $\alpha = 0.05$
- **Критическая область**: $z_{0.025} = 1.96$

$$K_L = 8 - 1.96 \cdot \frac{2.5}{\sqrt{50}} \approx 7.31, \quad K_R = 8 + 1.96 \cdot \frac{2.5}{\sqrt{50}} \approx 8.69$$

- **Решение**: $\bar{x} = 7.2 < K_L = 7.31$, поэтому отклоняем H_0.
- **Вывод**: Имеются достаточно статистически значимые основания для утверждения, что переделка сайта изменила среднюю длительность сессии. Консервативная гипотеза отклоняется в пользу альтернативной.

Гипотезы о математическом ожидании при неизвестной дисперсии

Предположения:

1. Случайная выборка $\mathcal{X} = \{X_1, \dots, X_n\}$ из неизвестного распределения с $\mu \equiv E[X_i]$, и $\sigma^2 \equiv \mathrm{Var}\,[X_i]$,
2. Истинную дисперсию считаем неизвестной,
3. Либо: большая выборка ($n > 30$), либо исследуемый случайный процесс имеет нормальное распределение

Гипотезы о математическом ожидании при неизвестной дисперсии

Пример 5: левосторонний тест

Производственная компания внедрила новый производственный процесс и хочет проверить, уменьшилось ли среднее время производства детали относительно целевого значения 30 минут. Более быстрое производство указывало бы на успешное внедрение новой системы.

- Исследуемая случайная величина: X — время производства. Дисперсия X неизвестна.
- Предполагается, что времена производства следуют нормальному распределению.

Данные:

- Выборка: $n = 16$ различных произведенных деталей
- Выборочное среднее: $\bar{x} = 27.5$ минут
- Выборочное стандартное отклонение: $s = 5$ минут

Гипотезы:

- $H_0 : \mu = 30$
- $H_1 : \mu < 30$

Гипотезы о математическом ожидании при неизвестной дисперсии

Левосторонний тест

Для тестирования $H_0 : E[X] = \mu = \mu_0$ против $H_1 : E[X] = \mu < \mu_0$:

- *Распределение при нулевой гипотезе*, когда мы безраздельно верим в то, что она верна:

$$\bar{X} \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right), \quad \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1\ df}$$

- Построение границы критической области теста относительно установленного уровня значимости:

$$P_{H_0}\left(\bar{X} < K\right) = \alpha$$

- Преобразование к t-распределению:

$$P_{H_0}(\bar{X} < K) = P_{H_0}\left(\frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} < \frac{K-\mu_0}{\frac{S}{\sqrt{n}}}\right) = P(t < -t_{(n-1,\alpha)})$$

- Восстановление границы отклоняющей области:

$$K = \mu_0 - t_{(n-1,\alpha)}\frac{S}{\sqrt{n}}$$

- Правило принятия решения: Отклонить H_0, если $\bar{x} < K$

Гипотезы о математическом ожидании при неизвестной дисперсии

Пример 5: решение

- **Гипотезы**: $H_0 : \mu = 30$ против $H_1 : \mu < 30$ (левосторонний тест)
- **Данные**: $\mu_0 = 30$, $s = 5$, $n = 16$, $\bar{x} = 27.5$, $\alpha = 0.05$
- **Критическая область**: $t_{15,0.05} = 1.753$

$$K = 30 - 1.753 \cdot \frac{5}{\sqrt{16}} = 27.81$$

- **Решение**: $\bar{x} = 27.5 < K = 27.81$, поэтому отклоняем H_0.
- **Вывод**: Имеются достаточно статистически значимые основания для утверждения, что новый производственный процесс уменьшил среднее время производства детали относительно целевых 30 минут. Консервативная гипотеза отклоняется в пользу альтернативной.

Гипотезы о доле бинарного признака

Иначе: о вероятности успеха в схеме Бернулли

Напоминание о выборочной доле

- Введём Y — случайную величину, показывающую количество положительных ответов в выборке, она имеет биномиальное распределение с $E[Y] = np$ и $Var[Y] = np(1-p)$.
- Тогда $\hat{p} = \frac{Y}{n}$ — случайная величина, называемая *выборочной долей* и показывающая долю положительных ответов к размеру выборки.
- Её свойства можно вывести из Y, а именно: $E[\hat{p}] = p$, $Var[\hat{p}] = \frac{p(1-p)}{n}$.
- Как следствие ИТМЛ, если $n > 30$, то:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right), \quad \text{или} \quad \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$$

Гипотезы о доле бинарного признака

Пример 6: правосторонний тест

Компания запустила новую телевизионную рекламную кампанию и хочет проверить, выше ли уровень узнаваемости бренда, чем было исторически ранее (40%).

- Выборка: $n = 200$ опрошенных клиентов
- Реализация выборочной доли: $\tilde{p} = 0.45$ (45% узнали бренд)

Гипотезы:

- $H_0 : p = 0.40$ (новая кампания не более эффективна)
- $H_1 : p > 0.40$ (новая кампания увеличивает узнаваемость бренда)

Гипотезы о доле бинарного признака

Правосторонний тест

Для тестирования $H_0 : p = p_0$ против $H_1 : p > p_0$:

- *Распределение при нулевой гипотезе*, когда мы безраздельно верим в то, что она верна:

$$\hat{p} \sim \mathcal{N}\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

- Построение критической (отклоняющей) области относительно уровня значимости:

$$P_{H_0}\left(\hat{p} > K\right) = \alpha$$

- Преобразование к стандартному нормальному распределению:

$$P_{H_0}\left(\hat{p} > K\right) = P_{H_0}\left(\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > \frac{K - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) = P(Z > z_\alpha)$$

- Восстановление границы отклоняющей области:

$$K = p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$$

- Правило принятия решения: Отклонить H_0, если $\tilde{p} > K$

Гипотезы о доле бинарного признака

Пример 6: решение

- **Гипотезы**: $H_0 : p = 0.40$ против $H_1 : p > 0.40$ (правосторонний тест)
- **Данные**: $p_0 = 0.40$, $n = 200$, $\tilde{p} = 0.45$, $\alpha = 0.05$
- **Критическая область**: $z_{0.05} = 1.645$

$$K = 0.40 + 1.645 \cdot \sqrt{\frac{0.40 \cdot 0.60}{200}} \approx 0.457$$

- **Решение**: $\tilde{p} = 0.45 < K = 0.457$, поэтому не отклоняем H_0.
- **Вывод**: Результаты тестирования не предоставляют достаточных статистически значимых оснований для отклонения нулевой гипотезы. Нет достаточных оснований утверждать, что новая рекламная кампания увеличила узнаваемость бренда выше 40%.