Проверка применимости критерия с помощью моделирования (равномерное распределение p-value). Как по распределению p-level сравнить критерии по мощности?

Мальцев А.С и Соколиков Е.А.

Санкт-Петербургский государственный университет
Прикладная математика и информатика

Санкт-Петербург
2019г.

Общий подход к проверке гипотез

- Гипотеза H_0 — некоторое предположение о случайной величине ξ $(\xi : (\Omega, \mathfrak{F}, P) \longrightarrow V)$;
- Выборка $x = (x_1, ..., x_n)$;
- Критерий в общем смысле:
  $V^n = \mathcal{A}^{\alpha}_{\text{дов}} \sqcup \mathcal{A}^{\alpha}_{\text{крит}}$ так, чтобы $P_{H_0}(x \in \mathcal{A}^{\alpha}_{\text{крит}}) = \alpha$, где α – уровень значимости, $0 < \alpha \leqslant 1$;
- Использование критерия:
  $x \in \mathcal{A}^{\alpha}_{\text{дов}} \Rightarrow H_0$ не отвергаем;
  $x \in \mathcal{A}^{\alpha}_{\text{крит}} \Rightarrow H_0$ отвергаем;
- α – вероятность отвергнуть гипотезу, если она верна.

Схема построения критериев

- Строим статистику $t = t(x_1, ..., x_n), t \in D \subset \mathbb{R}$:
  - распределение t, если верна H_0, известно;
  - t измеряет то, насколько выборка не соответствует гипотезе;
- $t \in D$ (например, $(-\infty, +\infty)$ или $[0, +\infty)$). Разбиваем $D = A^{\alpha}_{\text{дов}} \sqcup A^{\alpha}_{\text{крит}}$:
  - $P_{H_0}(t \in A^{\alpha}_{\text{крит}}) = \alpha$;
  - включаем в $A^{\alpha}_{\text{крит}}$ значения наиболее далекие от идеального (— соображения разумности выбора $A^{\alpha}_{\text{крит}}$).

Ошибки первого и второго рода

- Альтернативная гипотеза H_1 – это некоторое предположение о случайной величине ξ, которое не пересекается с гипотезой H_0;
- Вероятность ошибки первого рода $\alpha_{\mathbf{I}} = P_{H_0}(t \in A^{\alpha}_{\text{крит}}) = \alpha$;
- Вероятность ошибки второго рода $\alpha_{\mathbf{II}} = P_{H_1}(t \in A^{\alpha}_{\text{дов}}) = \alpha_{\mathbf{II}}(n, \alpha, H_1)$.

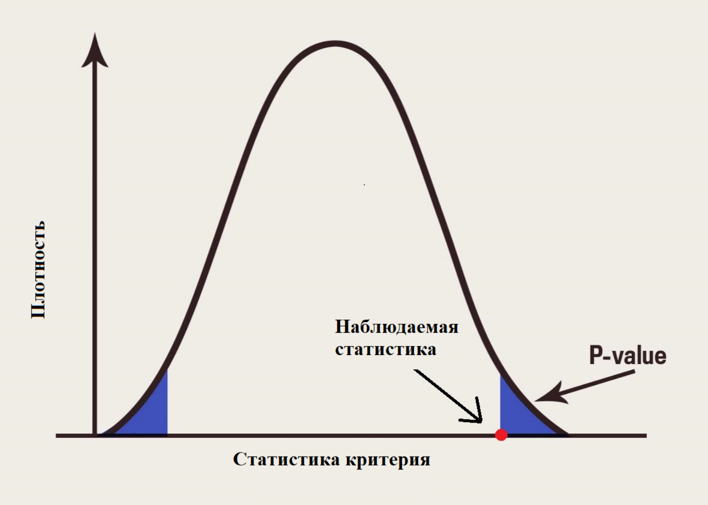
		Верная гипотеза	
		H0	H1
Результат применения критерия	H0	H0 верно не отвергнута	H0 неверно не отвергнута (Ошибка второго рода)
	H1	H0 неверно отвергнута (Ошибка первого рода)	H0 верно отвергнута

Виды критериев и мощность

- $\alpha_{\mathbf{I}} = \alpha$ — точный критерий;
- $\alpha_{\mathbf{I}} < \alpha$ — консервативный критерий;
- $\alpha_{\mathbf{I}} > \alpha$ — радикальный критерий;
- $\alpha_{\mathbf{I}} \underset{n\to\infty}{\longrightarrow} \alpha$ — асимптотический критерий;
- $\beta_n = \beta = 1 - \alpha_{\mathbf{II}} = \beta(n, \alpha, H_1)$ — мощность критерия против альтернативной гипотезы H_1;
- $\beta_n \underset{n\to\infty}{\longrightarrow} 1$ — состоятельность критерия против альтернативы H_1.

p-value

- p-value (p-level) – это максимальный уровень значимости, при котором гипотеза не отвергается.



Проверка применимости критерия

- P-value можно рассматривать, как случайную величину: $p = p(x_1, \dots, x_n)$;
- $\alpha_{\mathbf{I}} = P_{H_0}(H_0 \text{ отвергается}) = \alpha \Leftrightarrow P_{H_0}(\alpha > p) = \alpha \Leftrightarrow P_{H_0}(p < \alpha) = \alpha$;
- Таким образом, если H_0 верна, то p-value равномерно распределено на $[0, 1]$.

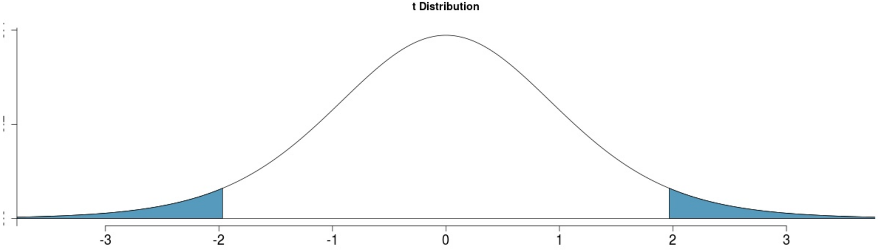
Алгоритм проверки применимости критерия:

- Моделируем m выборок $(x_1, \dots, x_n)$ размера n из распределения случайной величины ξ, при этом распределение ξ соотвествует нулевой гипотезе;
- Подсчитываем p-value для каждой выборки. Тогда получаем $(p_1, ..., p_m)$;
- Строим эмпирическую функцию распределения: $\hat{\mathcal{F}}_m(y) = \dfrac{\#\{p_i < y\}}{m}$.

Пример моделирования

- Модель: $\xi \sim \mathcal{N}(a, \sigma^2)$;
- Гипотеза: $H_0 : a = a_0 = 0$;
- В качестве критерия, применимость которого будем проверять, возьмем одновыборочный t-критерий Стьюдента (one sample t-test):

$$t = \frac{\sqrt{n}(\overline{x} - a_0)}{\tilde{s}} \sim t(n-1)$$



- Моделируем 20000 выборок размера 200 из распределения $\mathcal{N}(0, 1)$;
- Подсчитываем выборку p-value: $(p_1, \ldots, p_{20000})$ и строим функцию распределения.

Сравнение функций распределения

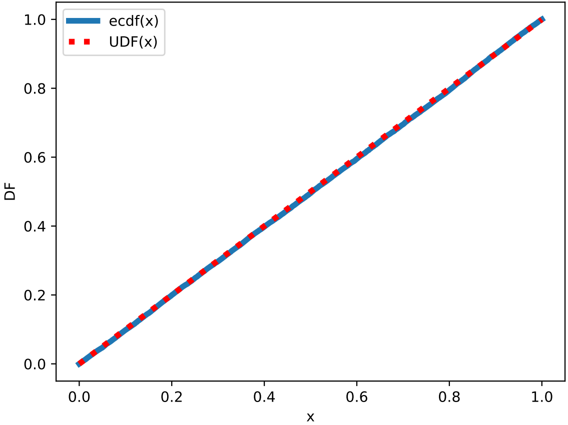
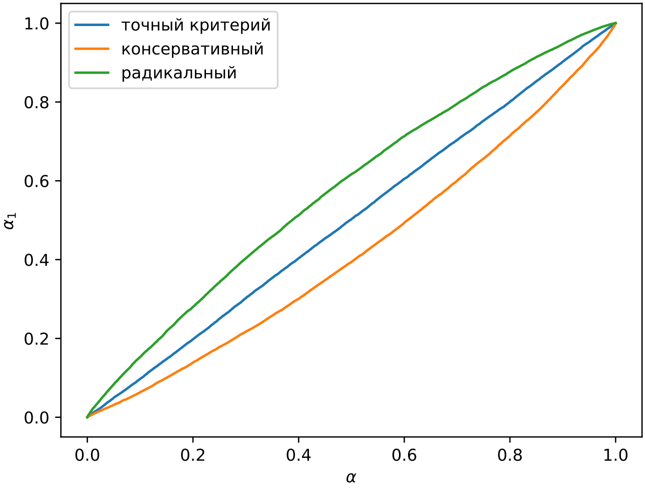


Рис.: Эмпирическая функция распределения p-levels.

Консервативные и радикальные критерии



Сравнение критериев по мощности

- $\alpha_{\mathbf{II}} = P_{H_1}(H_0 \text{ не отвергается}) \Leftrightarrow \alpha_{\mathbf{II}} = P_{H_1}(\alpha < p) \Leftrightarrow$
$\Leftrightarrow \alpha_{\mathbf{II}} = 1 - P_{H_1}(p < \alpha) \Leftrightarrow P_{H_1}(p < \alpha) = \beta$;
- Таким образом, если верна альтернативная гипотеза, то через эмпирическую функцию распределения p-value можно найти мощность критерия против альтернативы H_1.

Алгоритм сравнения критериев по мощности:

- Моделируем m выборок $(x_1, \ldots, x_n)$ размера n из распределения случайной величины ξ, при этом распределение ξ соотвествует альтернативной гипотезе;
- Подсчитываем p-value для каждой выборки;
- Оцениваем мощности.

Пример моделирования

- Модель: $\xi \sim \mathcal{N}(a, \sigma^2)$;
- Гипотеза: $H_0 : a = 0$;
- В качестве критерия возьмем одновыборочный t-критерий Стьюдента (one sample t-test);
- Альтернативные гипотезы: $H_1^{(1)} : a = 0.1$, $H_1^{(2)} : a = 0.3$;
- Моделируем 20000 выборок размера 200 из распределения $\mathcal{N}(0.1, 1)$ и столько же из распределения $\mathcal{N}(0.3, 1)$;
- Подсчитываем выборки p-value: $(p_1^{(1)}, \dots, p_{20000}^{(1)})$ и $(p_1^{(2)}, \dots, p_{20000}^{(2)})$ и строим функции распределения.

Сравнение функций распределения

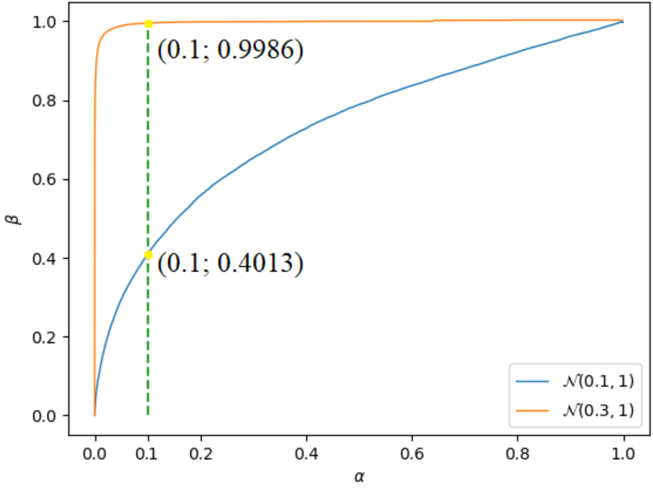


Рис.: Эмпирическая функция распределения p-levels.

Python code

```
EST = 0
VAR = 1
SAMPLE_SIZE = 200
PVAL_SAMPLE_SIZE = 20000
SIZE = (SAMPLE_SIZE, PVAL_SAMPLE_SIZE)


def generate_pval(mean):
  return stats.ttest_1samp(np.random.normal(
      mean, VAR, size=SIZE), EST)[1]

def first_graph():
  pval = generate_pval(0)
  //generate_pval(0.1) - power
  ecdf = ECDF(pval)
  plt.plot(ecdf.x, ecdf.y, linewidth=4, label='ecdf(x)')
  plt.show()
```

R code

```
samp <- matrix(rnorm(20000 * 200, 0, 1), nrow = 20000)
//rnorm(20000 * 200, 0.1, 1) - power
pval <- apply(samp, 2, function(x)
        t.test(x, mu = 0)$p.value)
distr <- ecdf(pval)
plot(distr)
```