Методы современной прикладной статистики
2. Основы проверки гипотез

Родионов Игорь Владимирович
vecsell@gmail.com

Весна, 2018

Постановка задачи

Пусть X (в нашем курсе, как правило, X – это выборка $(X_1, \ldots, X_n)$) – наблюдение из неизвестного распределения P_X. Основная задача – по наблюдению X сделать выводы о распределении P_X.

Предположим, что $P_X \in \mathcal{P}$, где $\mathcal{P}$ – некий класс распределений, которому заведомо принадлежит P_X.

Основные понятия

- **Основная гипотеза**: $H_0: \; P_X \in \mathcal{P}_0$, где $\mathcal{P}_0 \subset \mathcal{P}$.

  Пример. Пусть известно, что выборка имеет нормальное распределение, хотим проверить, что выборка распределена по закону $N(0,1)$. Тогда $\mathcal{P} = \{N(a,\sigma^2), a \in \mathbb{R}, \sigma^2 > 0\}$, а $\mathcal{P}_0 = \{N(0,1)\}$.

- **Альтернативная гипотеза (или альтернатива)**: $H_1 : P_X \in \mathcal{P}_1$, где $\mathcal{P}_1 \subset \mathcal{P} \backslash \mathcal{P}_0$.

- Гипотеза называется **простой**, если $\mathcal{P}_0$ (или $\mathcal{P}_1$) состоит из одного распределения.

- **Статистика критерия**: $T(X)$ – такая статистика, что при $P_X \in \mathcal{P}_0$ мы либо знаем её распределение, либо можем оценить сверху вероятности её редких значений.

- Если правило проверки гипотезы выглядит так:

$$\textit{если } T(X) \in S, \ \textit{то отвергнуть } H_0,$$

то S называется **критическим множеством**, а само правило называют **критерием**.

- Критерии бывают
    - двусторонние, $\{T(X) > u_{1-\alpha} \cup T(X) < u_\alpha\}$;
    - односторонние, которые делятся на правосторонние, $\{T(X) > u_{1-\alpha}\}$, и левосторонние, $\{T(X) < u_\alpha\}$;
    - более сложные.

- **Уровень значимости** критерия: такое α, что $P_0(T(X) \in S) \leq \alpha \ \forall P_0 \in \mathcal{P}_0$.

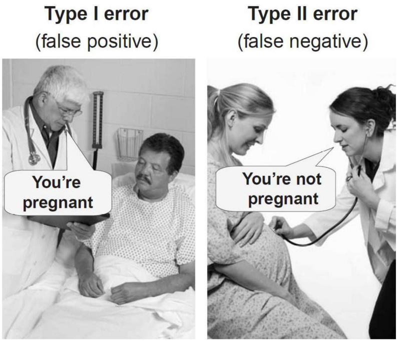
- **Размером** критерия называется его минимальный уровень значимости, т.е. такое α, что

$$\alpha = \sup_{P_0 \in \mathcal{P}_0} P_0(T(X) \in S).$$

Уровень значимости выбирается исследователем. Его обычные значения – $0.1, \ 0.05$ или 0.01.

Основные понятия

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода (False negative)
H_0 отвергается	Ошибка первого рода (False positive)	H_0 верно отвергнута

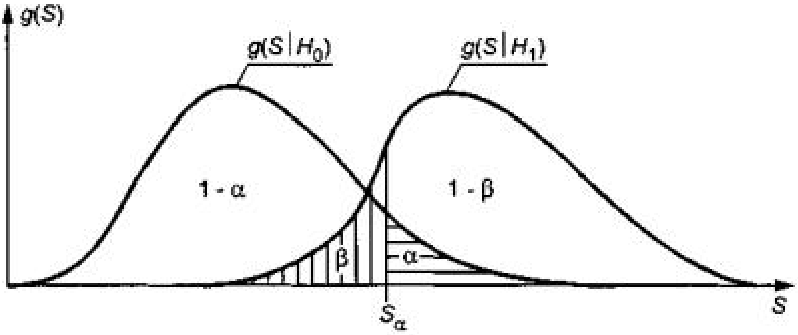


- **Функция мощности критерия**:
  $Q(S, P) = P(T(X) \in S)$.

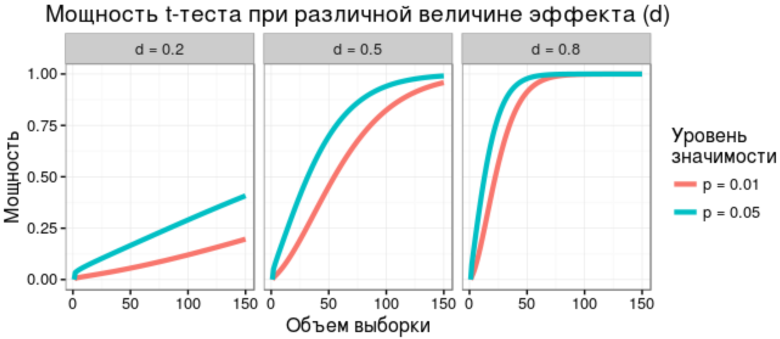
Тогда $Q(S, P_0)$, $P_0 \in \mathcal{P}_0$, – вероятность ошибки I рода на распределении P_0, а $1 - Q(S, P_1)$, $P_1 \in \mathcal{P}_1$, – вероятность ошибки II рода на распределении P_1.

Получаем, что, уменьшая вероятность ошибки I рода (т.е. уменьшая S), мы неизменно увеличиваем вероятность ошибки II рода, поэтому выбирать слишком низкий уровень значимости не рекомендуется.

Основные понятия



На рисунке – вероятности ошибок I и II родов в задаче различения 2 простых гипотез.



На рисунке – функции мощности t-критерия Стьюдента при различных уровнях значимости, если выборка сделана из распределения $N(d, \sigma^2)$.

- Критерий называется **состоятельным**, если $\forall P \in \mathcal{P}_1$ выполнено

$$P(T_n(X) \in S) \to 1 \text{ при } n \to \infty.$$

- Критерий называется **несмещенным**, если

$$\sup_{P \in \mathcal{P}_0} P(T(X) \in S) \leq \inf_{P \in \mathcal{P}_1} P(T(X) \in S).$$

- Критерий S **мощнее** критерия R, если уровни значимости этих критериев совпадают и $\forall P \in \mathcal{P}_1$ выполнено

$$P(T(X) \in S) \geq P(T(X) \in R).$$

- Критерий S называется **равномерно наиболее мощным** критерием (р.н.м.к.), если он мощнее любого другого критерия того же уровня значимости.

Р.н.м. критерии существуют далеко не во всех ситуациях. В задаче различения двух простых гипотез $H_0 : P = P_0$ против $H_1 : P = P_1$ р.н.м.к. существует всегда, как утверждает следующая лемма.

Лемма Неймана-Пирсона.
Пусть $S_\lambda = \{x : p_1(x) - \lambda p_0(x) \geq 0\}$, где p_i – плотность распределения P_i по мере μ, $i = 0, 1$. Пусть критерий R того же уровня значимости, что и критерий S_λ, т.е. $P_0(X \in R) \leq P_0(X \in S_\lambda)$. Тогда
1) $P_1(X \in R) \leq P_1(X \in S_\lambda)$ (т.е. S_λ мощнее R);
2) $P_1(X \in S_\lambda) \geq P_0(X \in S_\lambda)$ (т.е. S_λ несмещенный).

Пример. Пусть $X_1, \ldots, X_n$ – выборка из $Exp(\alpha)$. Построим р.н.м. критерий в задаче различения гипотезы $H_0 : \alpha = 2$ против альтернативы $H_1 : \alpha = 3$. Функция правдоподобия равна:

$$f(X, \alpha) = f(X_1, \ldots, X_n, \alpha) = \prod_{i=1}^{n} \alpha e^{-\alpha X_i} = \alpha^n e^{-\alpha \sum_i X_i},$$

тогда р.н.м. критерий

$$S_\lambda = \left\{ \frac{f(X,3)}{f(X,2)} \geq \lambda \right\} = \left\{ \frac{3^n}{2^n} e^{-\sum_i X_i} \geq \lambda \right\} = \left\{ \sum_{i=1}^{n} X_i \leq \lambda_1 \right\},$$

где $\lambda_1 = n \ln(2/3) + \ln \lambda$.

Далее, найдем такое λ, что критерий будет иметь уровень значимости γ. Если гипотеза H_0 верна, то $\sum_{i=1}^{n} X_i \sim \Gamma(n, 2)$. Отсюда, $\lambda_1 = u_\gamma$, где u_γ – γ-квантиль распределения $\Gamma(n, 2)$. Значит, $\lambda = (3/2)^n e^{u_\gamma}$.

Пример

Пусть $X_1, \ldots, X_n$ – выборка из $Pois(\lambda)$. Построим критерий проверки гипотезы $H_0 : \lambda = \lambda_0$ против альтернативы $H_1 : \lambda \neq \lambda_0$ с помощью ЦПТ. При выполнении гипотезы H_0,

$$\frac{\sum_{i=1}^n X_i - n\lambda_0}{\sqrt{n\lambda_0}} \xrightarrow{d} N(0,1), \ \ n \to \infty.$$

Тогда критерий будет выглядеть так:

$$\text{если } \left| \frac{\sum_{i=1}^n X_i - n\lambda_0}{\sqrt{n\lambda_0}} \right| > u_{1-\frac{\alpha}{2}}, \ \text{ то отвергать } H_0,$$

где $u_{1-\frac{\alpha}{2}}$ – $(1-\alpha/2)$-квантиль $N(0,1)$. В отличие от критерия Неймана-Пирсона, данный критерий является **асимптотическим**, т.е. его уровень значимости лишь стремится к α при $n \to \infty$, а не равен α в точности.

Критерий Вальда

Но дисперсия не всегда является функцией от известного параметра λ_0, как в предыдущем примере, поэтому в общем случае стоит заменять дисперсию на её оценку s^2 (сходимость к $N(0,1)$ сохранится ввиду леммы Слуцкого). Получаем критерий

$$\text{если } \left| \frac{\sum_{i=1}^{n} X_i - n\lambda_0}{\sqrt{ns^2}} \right| > u_{1-\frac{\alpha}{2}}, \text{ то отвергать } H_0,$$

который называется **критерием Вальда**.

Монотонное отношение правдоподобия

Говорят, что доминируемое семейство распределений $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ имеет монотонное отношение правдоподобий по статистике $T(X)$, если $\forall \theta_0 < \theta_1$ функция $\frac{f(X,\theta_1)}{f(X,\theta_0)}$ является монотонной функцией от $T(X)$, причем монотонность одна и та же $\forall \theta_0 < \theta_1$.

Теорема(о монотонном отношении правдоподобия)
Пусть в задаче проверки гипотезы $H_0 : \theta \le \theta_0$ (или $\theta = \theta_0$) против альтернативы $H_1 : \theta > \theta_0$ отношение $\frac{f(X,\theta_2)}{f(X,\theta_1)}$ возрастает по $T(X)$ при $\forall \theta_1 < \theta_2$. Если $\exists c_\alpha$ такое, что $P_{\theta_0}(T(X) \ge c_\alpha) = \alpha$, то $S = \{T(X) \ge c_\alpha\}$ – р.н.м.к. уровня значимости α для проверки H_0 против H_1.

На практике часто возникает задача различения двух семейств распределений. Для этих целей можно использовать критерий отношения правдоподобия (RML-тест).

Пусть $H_0 : \theta \in \Theta_0$ и $H_1 : \theta \in \Theta_1$, где $\Theta_0 \cap \Theta_1 = \varnothing$ и $\Theta_0 \cup \Theta_1 \subseteq \Theta$. Введем статистику

$$\lambda_n(X, \Theta_0) = \frac{\sup_{\theta \in \Theta_0} f(X, \theta)}{\sup_{\theta \in \Theta} f(X, \theta)},$$

также используется статистика

$$\lambda_n'(X) = \frac{\sup_{\theta \in \Theta_0} f(X, \theta)}{\sup_{\theta \in \Theta_1} f(X, \theta)}.$$

Теорема Уилкса.
Пусть $\Theta \subset \mathbb{R}^k$, $\dim(\Theta) = k$, $\dim(\Theta_0) = s < k$, тогда при верности гипотезы $H_0 : \theta \in \Theta_0$ выполнено

$$-2 \ln \lambda_n(X, \Theta_0) \xrightarrow{d} \chi^2_{k-s}, \;\; n \to \infty.$$

Кроме того, критерий $\widetilde{S} = \{x : -2 \ln \lambda_n(X, \Theta_0) \geq u_{1-\alpha}\}$, является состоятельным, где $u_{1-\alpha}$ – $(1-\alpha)$-квантиль распределения χ^2_{k-s}.

Пусть $H_0 : P \in \mathcal{P}_0 = \{P_\theta, \theta \in \Theta\}$ и $H_0 : P \in \mathcal{P}_1 = \{\widetilde{P}_\gamma, \gamma \in \Gamma\}$. Рассмотрим статистику

$$K = \frac{P(\mathcal{P}_0|X)}{P(\mathcal{P}_1|X)} = \frac{\int_\Theta f(X,\theta)q(\theta)d\theta}{\int_\Gamma \widetilde{f}(X,\gamma)\widetilde{q}(\gamma)d\gamma}.$$

Заметим, что распределение статистики K не зависит от параметров θ и γ, поэтому можно найти её распределение при условии верности H_0 и H_1 (или промоделировать его) и в соответствии с этим построить критерий.

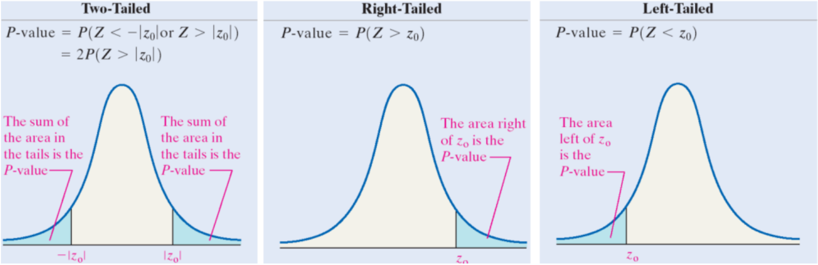
Но если нет времени, то есть шкала Джеффри!

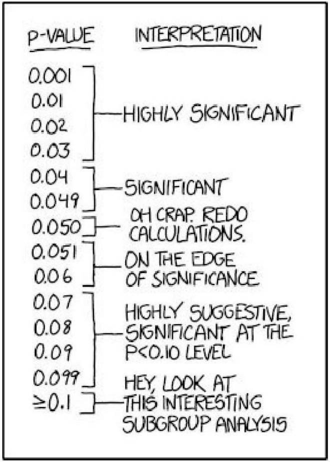
K	верна ли H_0?
1-3	нельзя определенно сказать
3-10	большие основания принять H_0
10-30	почти наверняка
>30	точно

Пусть значение статистики критерия $T(X)$ на наблюдении X равно t. Тогда p-значение – такая величина, которая является функцией от t и равна вероятности того, что $T(X)$ (на другой реализации наблюдения X) примет значение "экстремальнее", чем t.

Правило проверки гипотезы с помощью p-значения выглядит так: если $p < \alpha$, где α – уровень значимости критерия, то отвергаем основную гипотезу.

В случае левостороннего критерия $p = P(T(X) < t)$, в случае правостороннего критерия $p = P(T(X) > t)$, в случае двустороннего критерия $p = 2\min\{P(T(X) < t), P(T(X) > t)\}$.





Примерно так принимаются решения о значимости эффекта на основании p-значения (при основной гипотезе, что эффект незначим).

Пример (habrahabr.ru/company/stepic/blog/250527/). Допустим, мы хотим выяснить, существует ли связь между пристрастием к шутерам и агрессивностью у школьников. Для этого отобрали группу школьников, играющих в шутеры, и группу школьников, не играющих в компьютерные игры.

В качестве показателя агрессивности возьмём количество драк с участием конкретного школьника за месяц, в качестве основной гипотезы – что связи нет. Допустим, мы сравнили показатели 2 этих групп с помощью критерия хи-квадрат на уровне значимости 0.05 и получили p-значение, равное 0.04.

О чем говорит p-значение 0.04 в данном случае?

1. Компьютерные игры — причина агрессивного поведения с вероятностью 96%;
2. Вероятность того, что агрессивность и компьютерные игры не связаны, равна 0.04;
3. Если бы мы получили p-значение больше, чем 0.05, это означало бы, что агрессивность и компьютерные игры никак не связаны между собой;
4. Вероятность случайно получить такие различия равняется 0.04.
5. Ни один из вариантов не верен.

Ключевой вопрос: Допустим, при проверке некоторой гипотезы двумя критериями p-значение первого критерия оказалось меньше уровня значимости, а p-значение второго критерия больше. Как следует поступить: отвергнуть гипотезу или принять её?

Ключевой вопрос: Допустим, при проверке некоторой гипотезы двумя критериями р-значение первого критерия оказалось меньше уровня значимости, а р-значение второго критерия больше. Как следует поступить: отвергнуть гипотезу или принять её?

Ответ: Зависит от ситуации.

Table 5. The probability of correct selection (PCS) based on Monte Carlo (MC) simulations and also based on asymptotic results (AS) when the data are from log-normal distribution for different values of p.

	n					
	20	40	60	80	100	200
$p = 0.9$						
MC	0.731	0.814	0.869	0.913	0.931	0.987
AS	0.770	0.852	0.900	0.930	0.951	0.990
$p = 0.8$						
MC	0.693	0.782	0.846	0.877	0.923	0.974
AS	0.739	0.817	0.866	0.899	0.923	0.978
$p = 0.7$						
MC	0.689	0.761	0.810	0.848	0.878	0.948
AS	0.710	0.783	0.831	0.865	0.892	0.960
$p = 0.6$						
MC	0.662	0.728	0.749	0.784	0.826	0.911
AS	0.682	0.748	0.793	0.828	0.829	0.932
$p = 0.5$						
MC	0.617	0.665	0.686	0.749	0.750	0.842
AS	0.655	0.713	0.755	0.787	0.813	0.896
$p = 0.4$						
MC	0.568	0.666	0.663	0.690	0.715	0.786
AS	0.628	0.678	0.714	0.743	0.768	0.849
$p = 0.3$						
MC	0.598	0.647	0.676	0.683	0.717	0.785
AS	0.601	0.641	0.671	0.695	0.716	0.789

Таблица вероятностей правильного принятия решения с помощью RML-теста в задаче различения логномального и вейбулловского распределений в случае, если выборка имеет логнормальное распределение. p – доля минимальных членов вариационного ряда, взятых для построения теста.

	$n = 100$ $k = 10$	$n = 100$ $k = 20$	$n = 200$ $k = 20$	$n = 200$ $k = 50$	$n = 500$ $k = 50$	$n = 1000$ $k = 50$	$n = 5000$ $k = 50$
log-Pareto(2)	0.28	0.71	0.49	0.98	0.84	0.6	0.19
log-Pareto(1)	0.06	0.12	0.07	0.3	0.09	0.06	0.04
log-Gamma	0.37	0.58	0.62	0.89	0.94	0.97	0.99
Cauchy	0.51	0.8	0.8	0.99	0.99	1	1
Pareto(2)	0.92	1	1	1	1	1	1
LN(0,1)	0.91	0.98	1	1	1	1	1

Таблица эмпирических вероятностей ошибок I рода (первые 2 строчки) и эмпирических мощностей критерия различения распределений с супер-тяжелыми и тяжелыми хвостами. k – количество максимальных членов вариационного ряда, используемых для построения критерия. Количество реализаций $m = 10000$.

218 I. Fraga Alves et al. / Journal of Statistical Planning and Inference 139 (2009) 213 – 227

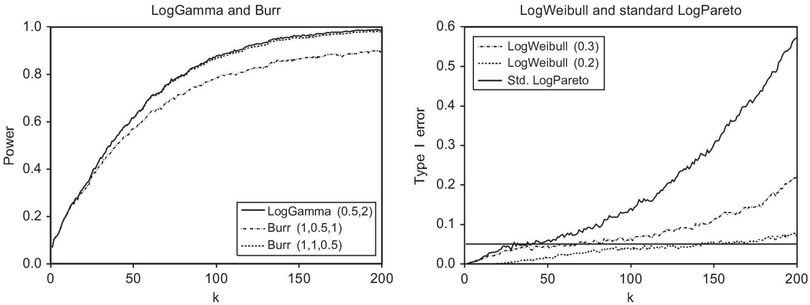


Fig. 1. Empirical power (left) and estimated type I error probability (right) of $T_n(k)$ at a nominal level $\alpha = 0.05$, built on 5000 samples of size $n = 1000$ from th prescribed parent distributions, all plotted against $k = 3, 4, \ldots, 200$.

Графики эмпирических вероятностей ошибок I рода (справа) и эмпирических мощностей (слева) того же критерия. Количество реализаций $m = 5000$, размер выборок $n = 1000$.

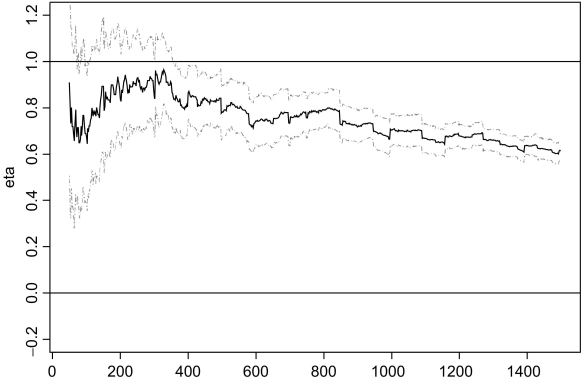
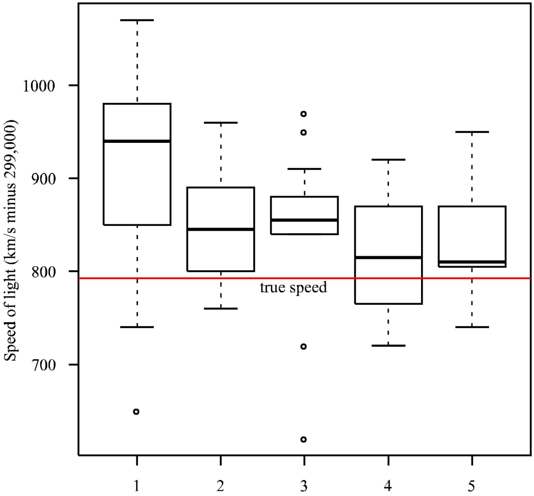


График значений некой оценки параметра θ и его 95%-доверительного интервала, построенного по этой оценке. Судя по этому графику, можно отвергнуть гипотезу о том, что $\theta = 1$, на уровне значимости 0.05.



Ящики с усами (boxplots) эксперимента Майкельсона-Морли по измерению скорости света.

Boxplot – графическое изображение некоторых свойств выборки или распределения. В частности, помогает сравнить между собой несколько наборов данный на предмет несовпадения их распределений.

1. Жирная линия в середине ящика - медиана выборки (или распределения, изредка - выборочное среднее или мат. ожидание);
2. Верхняя и нижняя границы ящика – верхняя и нижняя квартили выборки соответственно (т.е. квантили уровня 0.75 и 0.25);

3 Границы верхнего и нижнего “усов” ящика – как правило, максимальное и минимальное значение выборки без выбросов;

4 Кружки выше и ниже границы усов – выбросы, которые определяются либо как не принадлежащие интервалу

$$(Z_{0.25} - 1.5(Z_{0.75} - Z_{0.25}), Z_{0.75} + 1.5(Z_{0.75} - Z_{0.25}))$$

значения выборки, либо с помощью статистических тестов.

Finita!