

# Гипотезы сами себя не проверят

*Критерии согласия*

**1.** Согласно закону Бенфорда, первая цифра  $\xi_1$  случайного числа с десятичной записью  $\xi_1 \dots \xi_n$  из достаточно широко диапазона имеет распределение

$$\mathsf{P}(\xi_1 \leq d) = \log_{10}(d+1), \quad d \in \{1, \dots, 9\}.$$

Для выборки из стран мира (данные можно взять, например, отсюда) и уровня значимости 0.05 проверить гипотезу о том, что численность населения подчиняется закону Бенфорда.

**2.** Критерий  $\chi^2$  подходит для проверки не только большого уклонения от нулевой гипотезы, но и подозрительно точного соответствия ей. В этом случае можно взять критерий вида

$$R'_\alpha = \{\mathbf{x}: \chi^2(\mathbf{x}) \geq \chi^2_{k-1, 1-\alpha/2} \vee \chi^2(\mathbf{x}) \leq \chi^2_{k-1, \alpha/2}\}.$$

Критерий выше будет отклонять нулевую гипотезу и когда данные далеки от теории, и когда они слишком хорошо ей описываются. Найдите  $\text{pvalue}(\mathbf{X})$  для такого критерия.

**3.** Прочитайте доказательство леммы Неймана-Пирсона и предложите достаточное условие, при котором р.н.м.к. из этой леммы будет единственным с точностью до  $\mu$ -п.н., где  $\mu$  — мера, по которой берутся плотности  $\rho_0(x)$  и  $\rho_1(x)$  из условия.

**4.** Докажите, что в модели сдвига  $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$ ,  $\theta \in \mathbb{R}$ , не существует р.н.м.к. для проверки гипотезы  $H_0: \theta = 0$  vs.  $H_1: \theta \neq 0$ .

**5.** Дивергенцией Кульбака-Лейблера и дивергенцией- $\chi^2$  двух дискретных распределений  $\mathsf{P} = (p_1, \dots, p_k)$  и  $\mathsf{Q} = (q_1, \dots, q_k)$  называются соответственно величины

$$KL(\mathsf{P} \parallel \mathsf{Q}) = \sum_{i=1}^k p_i \cdot \log \frac{p_i}{q_i} \text{ и } \chi^2(\mathsf{P} \parallel \mathsf{Q}) = \sum_{i=1}^k \frac{(p_i - q_i)^2}{q_i}.$$

Пусть  $k$ -гранный кубик с вероятностями выпадения  $i$ -ой грани  $p_i^0$  подкидывают  $n$  раз,  $(\nu_1, \dots, \nu_k)$  — наблюдаемые частоты, а  $p_i = \nu_i/n$ . Докажите, что

$$\frac{\chi^2(\mathsf{P} \parallel \mathsf{P}^0)}{KL(\mathsf{P} \parallel \mathsf{P}^0)} \xrightarrow{d} 2,$$

откуда выведите, что

$$2n \cdot KL(\mathsf{P} \parallel \mathsf{P}^0) = \sum_{i=1}^k 2\nu_i \cdot \log \frac{\nu_i}{np_i^0} \xrightarrow{d} \chi^2_{k-1}, \quad n \rightarrow \infty.$$

Что можно сказать про сходимость статистики  $2n \cdot D(\mathsf{P}^0 \parallel \mathsf{P})$ ? Какую из них лучше использовать?

**6.\*** Модифицируйте критерий Колмогорова таким образом, чтобы, во-первых, в качестве основной гипотезы  $H_0: \mathsf{P} = \mathsf{P}_0$  можно было выбрать произвольное распределение  $\mathsf{P}_0$  (необязательно непрерывное), и, во-вторых, критерий остался состоятельным против альтернативы  $H_1: \mathsf{P} \neq \mathsf{P}_0$ .