Критерий Колмогорова-Смирнова

Н.П. Алексеева,

СПбГУ, математико-механический факультет

2020 г.

КС-тест

Изучается вопрос, насколько непрерывная функция распределения F(t) может отличаться от эмпирической $F_n(t)$.

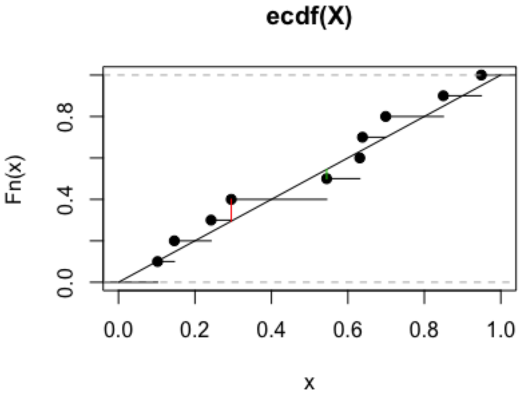
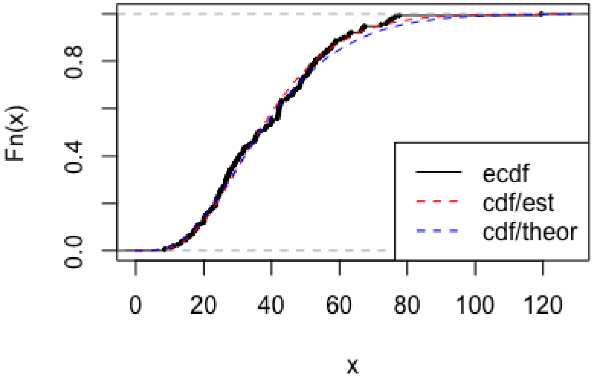


Figure:

КС-тест



Построение КС-теста

Исследуем сначала положительные отклонения $\mathrm{F} - \mathrm{F_n}$ и обозначим через $\Delta = \max(\mathrm{F} - \mathrm{F_n})$. Введем монотонное преобразование $\eta = \mathrm{F}(\xi)$, сохраняющее максимум Δ, что при $0 < \mathrm{t} < 1$

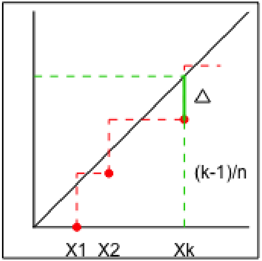
$$\mathrm{P}\{\eta < \mathrm{t}\} = \mathrm{P}\{\mathrm{F}(\xi) < \mathrm{t}\} = \mathrm{P}\{\xi < \mathrm{F}^{-1}(\mathrm{t})\} = \mathrm{F}(\mathrm{F}^{-1}(\mathrm{t})) = \mathrm{t}$$

События, при котором $\mathrm{t} < 0$ или $\mathrm{t} > 1$ невозможны.

$$\mathrm{F}(\mathrm{t}) = \begin{cases} 0, & \mathrm{t} \leq 0, \\ \mathrm{t}, & 0 < \mathrm{t} < 1, \\ 1, & \mathrm{t} \geq 1 \end{cases} .$$

В точке x_k функция $F_n(t)$ совершает скачок от $(k-1)\delta$ до $k\delta$, где $\delta = \frac{1}{n}$, $F_n(x_k - 0) = F_n(x_k) = \frac{k-1}{n}$, $F_n(x_k + 0) = \frac{k}{n}$.

Пусть $\Delta = x_k - (k-1)\delta$. Событие $\Delta > \epsilon$ наступает тогда, когда хотя бы одна из разностей $x_k - (k-1)\delta$ больше ϵ.



Так как все $x_1, \ldots, x_n$ независимы и равномерно распределены, то

$$Q = P\{\Delta > \epsilon\} = \int \ldots \int_G dx_1 \ldots dx_n,$$
$$G = \{0 < x_1 < 1, \ldots, 0 < x_n < 1, \Delta > \epsilon\},$$
$$Q = n!q, \quad q = P\{0 < x_1 < x_2 < \ldots < x_n < 1, \Delta > \epsilon\}.$$

Пусть q_k вероятность того, что это событие произошло в точке с индексом k и не произошло в точках с меньшими индексами ($q = q_1 + \ldots + q_n$).

$$0 < x_1 < \ldots < x_n < 1,$$
$$x_k - (k-1)\delta > \epsilon,$$
$$x_j - (j-1)\delta \le \epsilon \ \text{ для } \ j < k.$$

Пусть сначала $k = 1$ и q_1 вероятность того, что событие $\Delta > \epsilon$ произошло в точке x_1. Тогда мы имеем неравенства

$$x_1 < \ldots < x_n < 1, \quad x_1 > \epsilon,$$

то есть все x_i лежат в интервале $(\epsilon, 1)$,

$$q_1 n! = (1-\epsilon)^n \implies q_1 = \frac{1}{n!}(1-\epsilon)^n \,. \tag{1}$$

Пусть $k = h + 1 > 1$. Тогда $q_k = q_{h+1} = p_h r_h$, где p_h,r_h соответственно вероятности событий

$$\begin{gathered} 0 < x_1 < \ldots < x_h, \\ x_j \leq \epsilon + (j-1)\delta,\ j = 1, \ldots, h \end{gathered} \tag{2}$$

$$\begin{gathered} x_{h+1} < x_{h+2} < \ldots < x_n < 1, \\ x_{h+1} > \epsilon + h\delta. \end{gathered} \tag{3}$$

Вероятность события (3) вычисляется по аналогии с (1)

$$r_h = \frac{1}{(n-h)!}(1 - \epsilon - h\delta)^{n-h},\ 1 - \epsilon - h\delta \geq 0.$$

Вероятность p_h события

$$0 < x_1 < \ldots < x_h,$$
$$x_j \le \epsilon + (j-1)\delta,\ \ j = 1, \ldots, h$$

имеет вид

$$p_h = \int\limits_0^{\epsilon} dx_1 \int\limits_{x_1}^{\epsilon+\delta} dx_2 \ldots \int\limits_{x_{h-1}}^{\epsilon+(h-1)\delta} dx_h,$$

при $h = 1$ находим $p_1 = \epsilon$. По индукции можно показать, что

$$p_h = \frac{\epsilon}{h!}(\epsilon + h\delta)^{h-1}.$$

Таким образом,

$$q_k = q_{h+1} = p_h r_h = \left(\frac{\epsilon}{h!}(\epsilon + h\delta)^{h-1}\right) \cdot \left(\frac{1}{(n-h)!}(1-\epsilon-h\delta)^{n-h}\right) =$$
$$= \frac{\epsilon}{h!(n-h)!}(\epsilon + h\delta)^{h-1}(1-\epsilon-h\delta)^{n-h},$$

$$Q = n!(q_1 + \ldots + q_n) = \sum_{h=0}^{H} C_n^h \epsilon(\epsilon + h\delta)^{h-1}(1-\epsilon-h\delta)^{n-h},$$

$$1-\epsilon-h\delta \geq 0, \implies h \leq n(1-\epsilon), H = \lfloor n(1-\epsilon) \rfloor .$$

Приближенное решение имеет вид

$$Q = e^{-2n\epsilon^2}$$

Доказательство $p_h = \int\limits_0^{\epsilon} dx_1 \int\limits_{x_1}^{\epsilon+\delta} dx_2 \ldots \int\limits_{x_{h-1}}^{\epsilon+(h-1)\delta}$

Пусть верно $p_h = \frac{\epsilon}{h!}(\epsilon + h\delta)^{h-1}$. При $h = 1$ имеем $p_1 = \varepsilon$. Рассмотрим

$$p_{h+1} = \int\limits_0^{\epsilon} dx_1 \int\limits_{x_1}^{\epsilon+\delta} dx_2 \ldots \int\limits_{x_h}^{\epsilon+h\delta} dx_{h+1}.$$

Если ввести вместо $x_2, \ldots, x_{h+1}$ новые переменные $y_1, \ldots, y_h$ по формулам $x_j = x_1 + y_{j-1}$, то окажется, что

$$p_{h+1} = \int\limits_0^{\varepsilon} R dx_1, \quad \text{где} \quad R = \int\limits_0^{\epsilon+\delta-x_1} dy_1 \int\limits_{y_1}^{\epsilon+2\delta-x_1} dy_2 \ldots \int\limits_{y_{h-1}}^{\epsilon+h\delta-x_1} dy_h$$

Если положить $\varepsilon + \delta - x_1 = \varepsilon'$, то можно применить индукционный переход

$$R = \frac{\epsilon'}{h!}(\epsilon' + h\delta)^{h-1} = \frac{\varepsilon + \delta - x_1}{h!}(\varepsilon + (h+1)\delta - x_1)^{h-1},$$

$$p_{h+1} = \int\limits_0^{\varepsilon} \frac{\varepsilon + \delta - x_1}{h!}(\varepsilon + (h+1)\delta - x_1)^{h-1} dx_1$$

$$\text{Итак,}\quad p_{h+1} = \int\limits_0^{\varepsilon} \frac{\varepsilon+\delta-x_1}{h!}(\varepsilon+(h+1)\delta-x_1)^{h-1}dx_1$$

Сделаем замену $z = \varepsilon + (h+1)\delta - x_1$

$$p_{h+1} = \int\limits_{\varepsilon+(h+1)\delta}^{(h+1)\delta} \frac{z-h\delta}{h!} z^{h-1}(-dz) = \frac{1}{h!} \int\limits_{(h+1)\delta}^{\varepsilon+(h+1)\delta} (z^h - h\delta z^{h-1})dz$$

$$1) \quad \int\limits_{(h+1)\delta}^{\varepsilon+(h+1)\delta} z^h dz = \frac{(\varepsilon+\delta(h+1))^{h+1}}{h+1} - \frac{\delta^{h+1}(h+1)^{h+1}}{h+1},$$

$$2) \quad h\delta \int\limits_{(h+1)\delta}^{\varepsilon+(h+1)\delta} z^{h-1}dz \quad = \quad h\delta\left(\frac{(\varepsilon+\delta(h+1))^h}{h} - \frac{\delta^h(h+1)^h}{h}\right)$$

$$\text{отсюда}\ \frac{(\varepsilon+\delta(h+1))^{h+1}}{h+1} - \frac{h\delta(\varepsilon+\delta(h+1))^h}{h} =$$

$$= \frac{(\varepsilon+\delta(h+1))^h}{h+1}(\varepsilon+\delta(h+1)-\delta(h+1)) = \frac{\varepsilon(\varepsilon+\delta(h+1))^h}{h+1}.$$