Доказательство критерия

1) $\Rightarrow$

Пусть X - выборка (с дискретным распределением), T - достаточная статистика

$L(X,\theta) = \prod\limits_{i=1}^{n} f(X_i,\theta)$

$x = (x_1, \dots, x_n)^T$

$f(x_1,\theta) = \mathbb{P}_\theta(X_1 = x_1)$, где x_1 принимает конечное или счётное число значений

$L(X,\theta) = \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x | T(X) = t = T(x)) \cdot \mathbb{P}_\theta(T(X) = t)$

Первый множитель от θ не зависит (в силу достаточности статистики), обозначим его за $h(x)$. Второй множитель обозначим за $g(T,\theta)$. В итоге:

$L(X,\theta) = h(x) \cdot g(T,\theta)$

2) $\Leftarrow$

Пусть выполнен критерий факторизации, $L(X,\theta) = h(x) \cdot g(T(x),\theta)$

$\mathbb{P}_\theta(X = x | T(X) = t = T(x)) = \frac{\mathbb{P}_\theta(\{X=x\} \cap \{T(X)=t\})}{\mathbf{P}_\theta(T(X)=t)} = \frac{\mathbf{P}_\theta(X=x)}{\mathbf{P}_\theta(T(X)=t)} = \frac{\mathbf{P}_\theta(X=x)}{\sum_{y:T(y)=t} \mathbf{P}_\theta(X=y)} = \frac{h(x)g(t,\theta)}{\sum_{y:T(y)=t} h(y)g(t,\theta)} = \frac{h(x)}{\sum_{y:T(y)=t} h(y)}$, получили выражение, не зависящее от θ - QED.

Следствие: любая эффективная оценка является достаточной статистикой для параметра

Доказательство

Пусть T^* - эффективная оценка $\tau(\theta)$

$T^* - \tau(\theta) = a(\theta) \frac{\partial}{\partial\theta} \ln L(X,\theta)$

$\int\limits_{\theta_0}^{\theta} \frac{T^*(X) - \tau(\theta)}{a(\theta)} d\theta = \ln L(X,\theta)|_{\theta_0}^{\theta}$

Левую часть обозначим за $\widetilde{g}(T^*,\theta)$

$\widetilde{g}(T^*,\theta) = \ln L(X,\theta) - \ln L(X,\theta_0)$

$e^{\widetilde{g}(T^*,\theta)} \cdot L(X,\theta_0) = L(X,\theta)$, левая часть это $g(T^*,\theta)h(X)$

T^* будет достаточной статистикой по критерию - QED

Примеры достаточных статистик

1) $N(\theta,\sigma^2)$ - $\overline{X}$ является эффективной оценкой параметра $\theta \Rightarrow \overline{X}$ является достаточной статистикой

2) $Bi(1,\theta)$ аналогично $\overline{X} = \theta$ эффективная оценка $\Rightarrow \overline{X}$ - достаточная статистика

3) $R(0,\theta)$, $\theta > 0$

Проверяем $X_{(n)}$

$f(x_1,\theta) = \begin{cases} \frac{1}{\theta}, & x_1 \in [0,\theta] \\ 0, & x_1 \notin [0,\theta] \end{cases}$

$L(X,\theta) = \begin{cases} \frac{1}{\theta^n}, & X_{(n)} \in [0,\theta] \\ 0, & x_1 \notin [0,\theta] \end{cases}$

$L(X,\theta) = \frac{e(\theta - X_{(n)})}{\theta^n} = g(X_{(n)},\theta)$, где $e(y) = \begin{cases} 1, & y \geq 0 \\ 0, & y < 0 \end{cases}$ - ступенька Хевисайда

По критерию факторизации $T = X_{(n)}$ - достаточная статистика

4) $R(\theta_1,\theta_2)$, $\Theta = \{(\theta_1,\theta_2) | \theta_1 < \theta_2\}$

$f(x_1,\theta_1,\theta_2) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & x_1 \in [\theta_1,\theta_2] \\ 0, & x_1 \notin [\theta_1,\theta_2] \end{cases}$

$L(X,\theta) = \frac{e(\theta_2 - X_{(n)})e(X_{(1)} - \theta_1)}{(\theta_2 - \theta_1)^n}$

$T = (T_1,T_2) = (X_{(1)}, X_{(2)})$

5) $N(\theta_1,\theta_2^2)$

$f(x_1,\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi}\theta_2} \exp(-\frac{1}{2}\frac{(x_1-\theta_1)^2}{\theta_2^2})$

$L(X, \theta_1, \theta_2) = \frac{1}{(2\pi)^{n/2}\theta_2^n} \exp(-\frac{1}{2\theta_2^2} \sum\limits_{i=1}^{n} (X_i - \theta_1)^2) = \frac{1}{(2\pi)^{n/2}\theta_2^n} \exp(-\frac{1}{2\theta_2^2} \sum (X_i - \overline{X} + \overline{X} - \theta_1)^2) = \frac{1}{(2\pi)^{n/2}\theta_2^n} \exp(-\frac{1}{2\theta_2^2} nS^2 - \frac{1}{2\theta_2^2} n(\overline{X} - \theta_1)^2)$

$T = (\overline{X}, S^2)$ - достаточная статистика

Утверждение - Любое взаимно-однозначное преобразование достаточной статистики снова приводит к достаточной статистике

Доказательство

Пусть T - достаточная статистика

$\widetilde{T} = \phi(T)$, где ϕ - взаимно-однозначное преобразование

$L(X, \theta) = g(T(X), \theta)h(X)$ из критерия факторизации для T

$T = \phi^{-1}(\widetilde{T})$

$L(X, \theta) = g(\phi^{-1}(\widetilde{T}), \theta)h(X)$ - критерий факторизации для $\widetilde{T}$

Пример - $N(\theta_1, \theta_2^2)$

$T = (\overline{X}, S^2)$

$S^2 = \overline{X^2} - \overline{X}^2$, $\overline{X^2} = S^2 + \overline{X}^2$

$(\overline{X}, \overline{X^2})$ - тоже достаточная статистика

Для чего нужны достаточные статистики?

Теорема Рао-Блэкуэлла-Колмогорова

Пусть T_1 - несмещённая оценка $\tau(\theta)$ и T - достаточная статистика

Тогда существует статистика $T_2 = H(T)$ такая, что $D_\theta T_2 \leq D_\theta T_1$

Доказательство

Используем условное матожидание (которое изучим в следующем семестре)

Есть $E_\theta T_1 = \tau(\theta)$, рассмотрим $E_\theta(T_1|T = t) = m(t, \theta)$ - это и есть условное матожидание

Если распределение T дискретно, то у него есть условное распределение, по которому можно интегрировать и получать условное матожидание

$E_\theta(T_1|T = t) = \int T_1(x) f_\theta(x|T = t)dx = m(t, \theta) = m(t)$, так как f_θ не зависит от параметра, T - достаточная статистика

Самое главное свойство условного матожидания:

$E_\theta(E_\theta(T_1|T)) = E_\theta T_1 = \tau(\theta)$, T_2 из условия теоремы это $E_\theta(T_1|T) = m(T)$

$D_\theta T_1 = E_\theta(T_1 - \tau(\theta))^2 = E_\theta(T_1 - T_2 + T_2 - \tau(\theta))^2 = E_\theta(T_1 - T_2)^2 + E_\theta(T_2 - \tau(\theta))^2 + 2E_\theta(T_1 - T_2)(T_2 - \tau(\theta))$

Докажем, что третье слагаемое, ковариация, равна нулю

$E_\theta(T_1 - T_2)(T_2 - \tau(\theta)) = E_\theta E_\theta((T_1 - T_2)(T_2 - \tau(\theta))|T) = E_\theta(T_2 - \tau(\theta))(E_\theta(T_1 - T_2|T)) = E_\theta(T_2 - \tau(\theta))(E_\theta(T_1|T) - E_\theta T_2) E_\theta(T_2 - \tau(\theta))(E_\theta T_2 - E_\theta T_2) = 0$

$D_\theta T_1 = E_\theta(T_1 - T_2)^2 + D_\theta T_2 \geq D_\theta T_2$ - равенство достигается, когда T_1 является достаточной статистикой - QED

Из теоремы можно сделать следующие выводы:

1) Если существует хотя бы одна несмещённая оценка функции $\tau(\theta)$, то существует как правило лучшая несмещённая оценка, являющаяся функцией от достаточной статистики. А значит оптимальная оценка является функцией достаточной статистики

2) Для того, чтобы найти оптимальную оценку, достаточно решить уравнение (1) относительно H

(1) $E_\theta H(T) = \tau(\theta)$, где T - достаточная статистика

3) Если T имеет плотность $p(t, \theta)$, то уравнение (1) запишется в виде (2)

(2) $\int H(t)p(t, \theta)dt = \tau(\theta)$ - уравнение несмещённости

Интегрирование ведётся по прямой или плоскости, в зависимости от размерности T

Если решение есть, то нашли кандидатов в оптимальные оценки

Если решений нет, то по теореме Рао-Блэкуэлла-Колмогорова оптимальной оценки нет

Если уравнение (2) имеет единственное решение, то $H(T)$ будет оптимальной оценкой функции $\tau(\theta)$

Критерием единственности решения (но не существования) является свойство полноты достаточной статистики

Достаточная статистика T называется полной, если для любой борелевской функции $\psi(t)$, определённой на множестве её значений, равенство $E_\theta \psi(T) = 0$ влечёт $\psi(t) = 0$ для (почти) всех значений T

$E_\theta(H_1(T) - H_2(T)) = 0$

$\psi(t) = H_1(t) - H_2(t) = 0 \ \Rightarrow \ H_1 = H_2$