

Доверителни интервали

Постановка.

- Случайна величина X ;
- $F_X(x, \theta)$ е разпределение, което искаме да разберем, като знаем че то зависи параметрично от някакъв параметър θ ;
- $\theta = (\theta_1, \dots, \theta_s)$ – имаме s на брой параметри (например при нормалното разпределение са $s = 2$: средно μ и стандартно отклонение σ)

$\vec{X}(X_1, \dots, X_n)$ – вектор от n независими еднакво разпределени наблюдения над X (прототипи на X). На база на тези наблюдения, които в крайна сметка ще бъдат сведени до някакви числа (за модела/за експеримента) трябва да намерим някаква оценка $\hat{\theta} = \hat{\theta}(\vec{X})$, която да я вземем близо до θ , така че да определи това
 точкова оценка
 разпределение $F_X(x, \theta)$.

Проблема на точковата оценка е, че сама по себе си тя е доста динамична. Това е логично, тъй като извадките могат да бъдат различни. Хубаво ще е освен тази $\hat{\theta}$, да имаме и някаква вероятност, с която истинския параметър θ да попада в интервал, който може да наречем *доверителен*.

Ще разглеждаме само едномерни параметри θ ($\theta \in \mathbb{R}$).

Цел. Ще търсим две числа $I_1 = I_1(\vec{X}) < I_2 = I_2(\vec{X})$ такава, за които $\mathbb{P}(I_1 < \theta < I_2) = \gamma$ (като γ обикновено е число по-голямо от 0.9 и по-малко от 0.999 и има следния смисъл: колкото по-малко е γ , толкова по-широки интервали ще се получават, за да може с по-голяма вероятност да хванем истинския параметър θ . Стандартно $\gamma = 0.95$, а $\gamma = 0.90$ е за не чак толкова важни изследвания. За медицински цели се използва $\gamma \geq 0.999$.)

Дефиниция (Централна статистика ЦС). Казваме, че $T = T(\vec{X}, \theta)$ е централна статистика, ако:

- 1) T е монотонна по θ
- 2) $\mathbb{P}(T < x) = F_T(x)$ не зависи от θ (T е функция на θ , но разпределението и не зависи от θ)

\oplus Имаме \vec{X} (вектор от наблюдения) и искаме да намерим някакъв доверителен интервал: (I_1, I_2) за θ , т.е. $\theta \in (I_1, I_2)$. Целта ни е да имаме някакво ниво на доверие $\gamma = \mathbb{P}(q_1 < T < q_2)$. За улеснение ще допуснем, че T расте по θ (но тя може и да намалява по θ).

Тъй като T е монотонна по θ , то

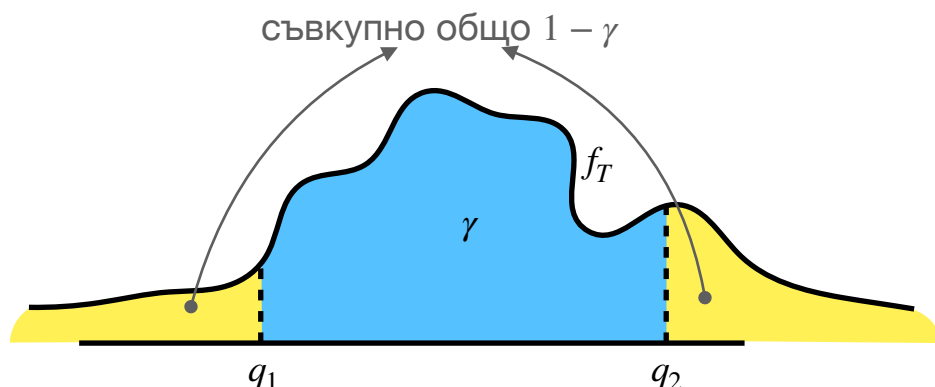
при фиксиран
вектор на

$$\gamma = \mathbb{P}(q_1 < T < q_2) \stackrel{\text{наблюдения } \vec{X}}{=} \mathbb{P}(T^{-1}(q_1) < \theta < T^{-1}(q_2)),$$

защото сме допуснали, че T расте по θ . Ако T намаляваше по θ , щяхме да имаме $\gamma = \mathbb{P}(T^{-1}(q_2) < \theta < T^{-1}(q_1))$. Следователно интервала ще е

$$I_1 = T^{-1}(q_1), I_2 = T^{-1}(q_2).$$

f_T - плътността на централната статистика (ЦС)

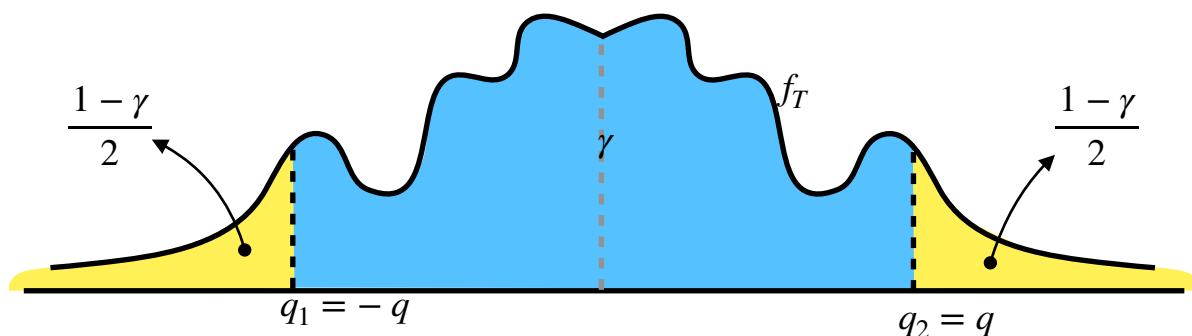


Има много начини, по които може да изберем q_1 и q_2 , така че вероятността между тях да е γ .

Имаме параметър $\theta \in \mathbb{R}$, който искаме да оценяваме. Търсим I_1 и I_2 , които да зависят от наблюденията \vec{X} и $I_1(\vec{X}) < I_2(\vec{X})$. Те образуват т.нар. *доверителен интервал* за θ с ниво на доверие $\gamma = \mathbb{P}(\theta \in (I_1, I_2))$.

T е централна статистика за θ , ако удовлетворява дефиницията за ЦС.

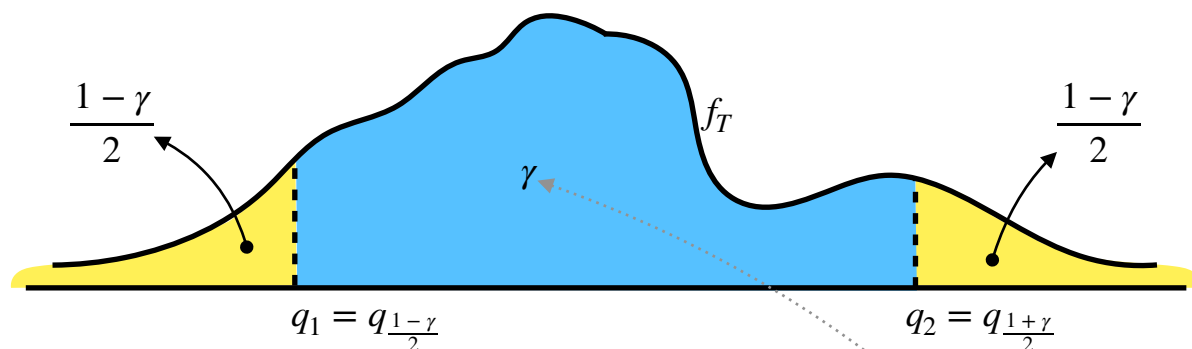
⊕ Ако T е симетрична случайна величина,



тогава се търсят такива q_1 и q_2 , че $q_1 = -q = -q_2$, за които

$$\mathbb{P}(T < -q) = 1 - \mathbb{P}(T < 1) = \frac{1 - \gamma}{2} \text{ (за да имаме по средата вероятност } \gamma)$$

По-общо, ако имаме някакво несиметрично разпределение на T :



$$\text{проверка : } \mathbb{P}(T < q_2) - \mathbb{P}(T < q_1) = \frac{1+\gamma}{2} - \frac{1-\gamma}{2} = \frac{1+\gamma-1+\gamma}{2} = \gamma$$

Имаме много начини, по които може да изберем q_1 и q_2 , но тези които демонстрираме по-горе са изпитани от практиката рецепти за избиране и имат конкретен смисъл за симетричните разпределения.

$$\underbrace{\gamma}_{\text{фиксирана}} = \mathbb{P}(q_1 < T < q_2) = \underbrace{\mathbb{P}(T^{-1}(q_1) < \theta < T^{-1}(q_2))}_{T \text{ нарастваща}}.$$

Тази вероятност γ е фиксирана и се задава предварително от изследователите. Какво може да оптимизираме ние като математици? - може да търсим q_1 и q_2 такива, за които е изпълнено:

$$\min_{\substack{q_1 < q_2 \\ \gamma = \mathbb{P}(q_1 < T < q_2)}} = \underbrace{\left\{ |T^{-1}(q_2) - T^{-1}(q_1)| \right\}}_{\substack{\text{искаме най-малък} \\ \text{интервал, за да може} \\ \text{да свием опциите за} \\ \theta \text{-максимално}}}.$$

Т.е. минимизираме доверителния интервал при фиксирано ниво на доверие!

\oplus $\underbrace{X \in \mathcal{N}(\mu, \sigma^2)}_{\substack{\text{случайна величина, която} \\ \text{искаме да изучаваме}}}$, σ^2 е известно, т.е. интересуваме се само от параметъра $\mu = \theta$ (едномерен). Искане да видим как може да оценим μ и да намерим за него доверителен интервал.

Ние знаем, че $\hat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$. Тогава $T(\vec{X}, \mu) = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \in \mathcal{N}(0,1)$, σ е известно

число. Това е така, защото
линейност

$$\sum_{j=1}^n X_j \stackrel{\substack{\text{на } \mathcal{N} \\ \text{и незав.}}}{\in} \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \frac{1}{n} \sum_{j=1}^n X_j \in \mathcal{N}\left(\frac{n\mu}{n}, \frac{n\sigma^2}{n^2}\right) = \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

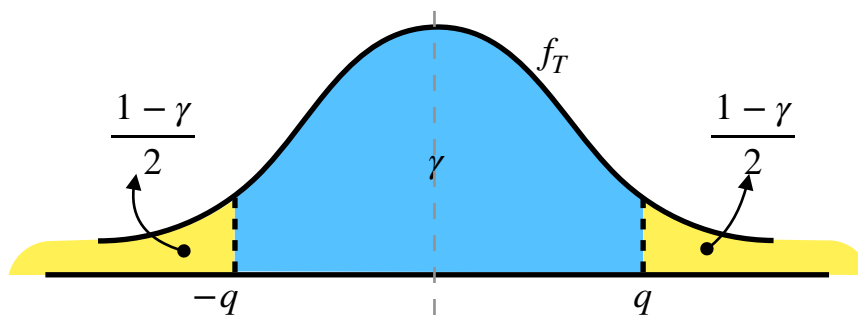
По-подробно обяснение на свойството линейност на нормалното разпределение:

$$\bar{X}_n = \frac{\sum_{j=1}^n X_j}{n} \in \mathcal{N}(x, y^2); \mathbb{E}\bar{X}_n = \frac{n\mu}{n} = x; \text{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{j=1}^n \sigma^2 = \frac{\sigma^2}{n} = y^2$$

T е намаляваща функция по μ и $\mathbb{P}(T \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy$ - не зависи от $\mu \stackrel{\text{def.}}{\Rightarrow}$

T е централна статистика за μ (тя е монотонна и намаляваща по μ и нейното разпределение съвпада с $\mathcal{N}(0,1)$, т.е. не зависи от θ)

Тогава, $\gamma = \mathbb{P}(-q < T < q)$, тъй като $\mathcal{N}(0,1)$ е симетрично:



Това ни гарантира, че $(-q, q)$ ще е най-тесния интервал, тъй като в $(-\infty, -q)$ и (q, ∞) е сбита най-малко маса (навсякъде извън тези интервали е сбита повече маса)

$q = q_{\frac{1}{2} + \frac{\gamma}{2}}$, който квантил го има в таблицата за нормалното стандартно разпределение.

$$\begin{aligned} \gamma &= \mathbb{P}(-q < T < q) = \mathbb{P}\left(-q_{\frac{1}{2} + \frac{\gamma}{2}} < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < q_{\frac{1}{2} + \frac{\gamma}{2}}\right) \Rightarrow \\ &\Rightarrow \mathbb{P}\left(\mu \in \underbrace{\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} \times q_{\frac{1}{2} + \frac{\gamma}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} \times q_{\frac{1}{2} + \frac{\gamma}{2}}\right)}_{I_1}\right) = \gamma \Rightarrow \\ &\Rightarrow I_1 = \bar{X}_n - \frac{\sigma}{\sqrt{n}} \times q_{\frac{1}{2} + \frac{\gamma}{2}}; I_2 = \bar{X}_n + \frac{\sigma}{\sqrt{n}} \times q_{\frac{1}{2} + \frac{\gamma}{2}}. \end{aligned}$$

$\oplus \quad X \in \mathcal{N}(\mu, \sigma^2)$, но този път не знаем σ .

Как да конструираме доверителен интервал само за μ ? Припомняме, че

$\hat{\mu} = \overline{X}_n$ и оценката за дисперсията е $s^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2$ и е неизместена.

независимо
дали знаем
или не σ

Фактора $\frac{1}{n-1}$ го има, тъй като тя е неизместена оценка за дисперсията.

Твърдение. Имаме, че $X \in \mathcal{N}(\mu, \sigma^2)$ и $\vec{X} = (X_1, \dots, X_n)$ са наблюдения над X . Тогава е вярно, че:

а) $\hat{\mu}$ е независимо от s^2 : $\hat{\mu} \perp s^2$;

б) $(n-1) \frac{s^2}{\sigma^2} \in \mathcal{X}^2(n-1)$.

$T = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$. Ако знаехме σ , последното щеше да е разпределено като $\mathcal{N}(0,1)$

неизвестно

(както направихме в предходния пример), но ние не знаем σ . Но друго, което знаем е, че:

$T = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{n-1}{n-1} \cdot \frac{s^2}{\sigma^2}}} = \frac{\overline{X}_n - \mu}{\frac{s}{\sqrt{n}}}$ е централна статистика!

$\in \mathcal{N}(0,1)$

$\in \mathcal{X}^2(n-1)$

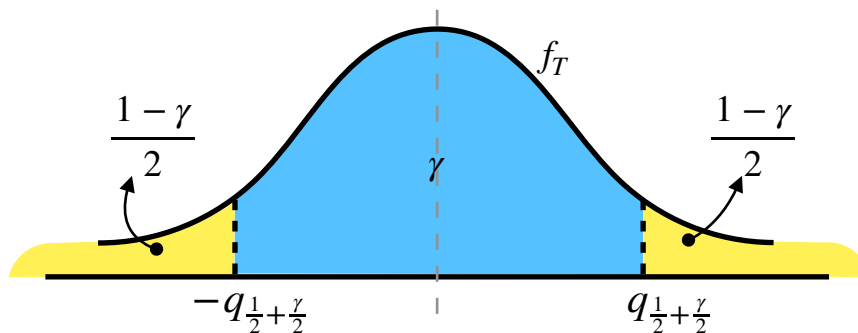
Тъй като $\frac{\overline{X}_n - \mu}{\frac{s}{\sqrt{n}}} = \frac{Z}{\sqrt{\frac{Y}{n-1}}}$, където $Z \in \mathcal{N}(0,1)$, $Y \in \mathcal{X}^2(n-1)$ и $Z \perp Y$.

Заключения.

- T е намаляваща по μ
- $T \in t(n-1)$ и не зависи от $\mu \Rightarrow T$ е централна статистика за μ .

$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}}; \quad Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}; \quad Y = \frac{(n-1)s^2}{\sigma^2}.$$

T е симетрична, тъй като в числителя имаме симетрична случайна величина (Z) \Rightarrow



$$\Rightarrow \mathbb{P}(-q_{\frac{1}{2}+\frac{\gamma}{2}} < T < q_{\frac{1}{2}+\frac{\gamma}{2}}) = \mathbb{P}\left(\mu \in \left(\bar{X}_n - q_{\frac{1}{2}+\frac{\gamma}{2}} \times \frac{s}{\sqrt{n}}, \bar{X}_n + q_{\frac{1}{2}+\frac{\gamma}{2}} \times \frac{s}{\sqrt{n}}\right)\right),$$

т.е. имаме, че I_1 и I_2 зависят от s и квантилите не са от нормалното разпределение а от t – Student's разпределението.

Бележки. $T - \frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} = \frac{Z}{\sqrt{\frac{Y}{n-1}}}, Z \perp Y, Z \in \mathcal{N}(0,1), Y \in \mathcal{X}^2(n-1)$

– $Y = \sum_{j=1}^{n-1} V_j$, където $(V_j)_{j=1}^{n-1}$ са независими и еднакво разпределени (i.i.d.) с $\mathcal{X}^2(1)$.

От ЗГЧ: $\frac{Y}{n-1} \xrightarrow[n \rightarrow \infty]{\text{п.с.}} \mathbb{E}V_1 = 1$. Следователно за големи n : $T \approx \frac{Z}{1} \approx \mathcal{N}(0,1)$;

– Ако знаем дисперсията: $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0,1)$ за произволни X_1, \dots, X_n, \dots

$\mathbb{E}X_1 = \mu; \text{Var}(X_1) = \sigma^2$.

\oplus $X \in \mathcal{N}(\mu, \sigma^2)$, знаем μ , $\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \mu)^2$. Искаме да оценим дисперсията.

Трябва да си конструираме централна статистика. ЦС не трябва да зависи от σ , но в $\hat{\sigma}^2$, X_j зависи от σ и трябва да отстраним тази зависимост.

Нагаждаме: $\frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{j=1}^n \underbrace{\left(\frac{X_j - \mu}{\sigma}\right)^2}_{Z_j^2} = \sum_{j=1}^n Z_j^2$, но $Z_j \in \mathcal{N}(0,1)$, тъй като центрираме и

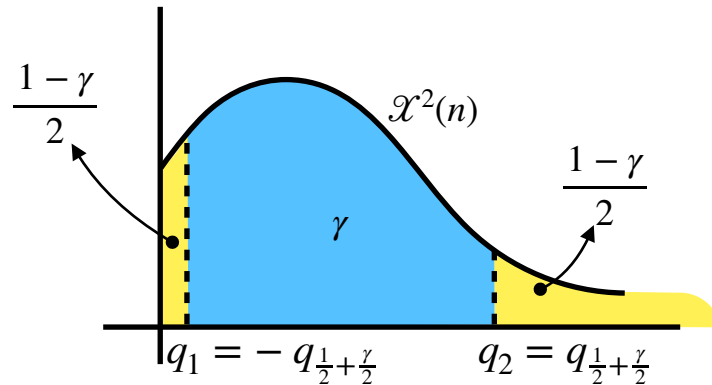
нормираме с неизвестна дисперсия. Следователно $\sum_{j=1}^n Z_j^2 \in \mathcal{X}^2(n)$, защото сумира

n квадрати на независими нормални стандартно разпределени $Z \Rightarrow$ статистиката

$T = \frac{n\hat{\sigma}^2}{\sigma^2}$ е централна по дефиниция (разпределението и е $\chi^2(n)$ – не зависи от σ^2

и тя е монотонно намаляваща по σ).

За квантилите:



$$\gamma = \mathbb{P}(q_1 < T < q_2) = \mathbb{P}\left(\frac{n\hat{\sigma}^2}{q_2} < \frac{n\hat{\sigma}^2}{q_1}\right) \Rightarrow \begin{cases} I_1 = \frac{n\hat{\sigma}^2}{q_2} \\ I_2 = \frac{n\hat{\sigma}^2}{q_1} \end{cases}$$

\oplus $X \in \mathcal{N}(\mu, \sigma^2)$, не знаем μ и се интересуваме от σ^2 . Знаем, че s^2 е неизместена оценка за σ^2 :

$(n-1)s^2 = \sum_{j=1}^n (X_j - \bar{X}_n)^2$ и нашата цел е да конструираме статистика за σ^2 , т.е. да

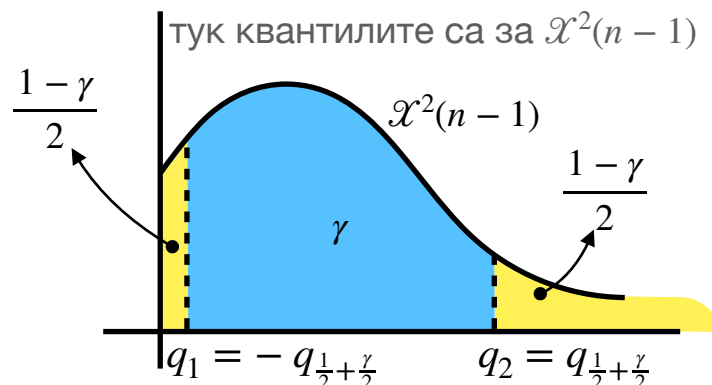
намерим ЦС, отговаряща на дефиницията. Ако разделим на σ^2 , от дясно не можем

да направим никакво заключение: $\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{j=1}^n (X_j - \bar{X}_n)^2}{\sigma^2}$, но знаем от

твърдението, че $T = \frac{(n-1)s^2}{\sigma^2} \in \chi^2(n-1)$ (с една степен по-малко, защото сме

изхабили една степен за оценката на μ). От n наблюдения все едно имаме $n-1$ наблюдения.

$\Rightarrow T$ е централна статистика (монотонна намаляваща по σ^2 и независима от σ^2)



$$\gamma = \mathbb{P} \left(q_1 < \frac{(n-1)s^2}{\sigma^2} < q_2 \right) = \mathbb{P} \left(\frac{(n-1)s^2}{q_2} < \sigma^2 < \frac{(n-1)s^2}{q_1} \right)$$

$$I_1 = \frac{1}{q_2}(n-1)s^2; \quad I_2 = \frac{1}{q_1}(n-1)s^2$$

При голямо n ($n > 30$ например) може да ползваме както от предходния пример - все едно знаем σ .

Проверка на хипотези

Теорията за тестване на хипотези е въведена и развита за първи път от Нейман и Пиърсън през 40-те години на миналия век. Тя най-често се задава със следната математическа постановка:

X е случайна величина с функция на разпределение $F_X(x, \theta)$, нулева хипотеза H_0 и алтернативна такава H_1 , където θ е оценявания параметър.

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

Искаме да конструираме някакво множество $W \in \mathbb{R}^n$ такова, че ако \vec{X} попадне в W ($\vec{X} \in W$), тогава отхвърляме H_0 (в полза на H_1), ако вектора от наблюдения \vec{X} не попадне в W , то тогава не отхвърляме H_1 .

H_0 и H_1 се наричат прости хипотези (нулева и алтернативна). Прости хипотези са $\theta = \theta_1$ (число). Сложни хипотези са $\theta > \theta_i$, $\theta \neq \theta_i$, $\theta \in I$ и т.н.

Цел. При наблюденията $\vec{X} \in \mathbb{R}^n$, търсим да конструираме $W \subseteq \mathbb{R}^n$: ако $\vec{X} \in W$, то отхвърляме H_0 и приемаме H_1 , а ако $\vec{X} \in \bar{W}$, то приемаме H_0 .

$$W \subseteq \mathbb{R}^n : \begin{cases} \vec{X} \in W \Rightarrow \text{отхвърляме } H_0 \\ \vec{X} \in \bar{W} \Rightarrow \text{приемаме } H_0 \end{cases}$$

Грешки, които може да допуснем:

- Грешка от $I^{-\text{ви}}$ род: Да отхвърлим H_0 , когато H_0 е вярна, т.е.
 $\alpha = \mathbb{P}(\vec{X} \in W | H_0)$
- Грешка от $II^{-\text{ри}}$ род: При положение, че е вярна хипотезата H_1 , ние сме приели H_0 , т.е. $\beta = \mathbb{P}(\vec{X} \in \bar{W} | H_1)$

$\pi = 1 - \beta$ се нарича *мощност* на W .

$$\oplus \quad \begin{aligned} H_0 &: \text{дадена ваксина е вредна } (\theta = \theta_0) \\ H_1 &: \text{ваксината не е вредна } (\theta = \theta_1) \end{aligned}$$

Има по-голям резон за H_0 да вземем по-опасната/рисковата хипотеза. Това е тази хипотеза, която е по-вероятно да я отхвърлим. Това е така, защото грешката от $I^{-\text{ви}}$ род ще бъде контролирана/задавана от изследователя (той ще казва дали иска/допуска да е 0.01, 0.05, 0.1 и т.н.)

Дефиниция (Оптимална критична област). При фиксирана грешка от $I^{-\text{ви}}$ род α , $W^* \subseteq \mathbb{R}^n$ се нарича оптимална критична област (ОКО), ако

$$\mathbb{P}(\vec{X} \notin W^* | H_1) = \min_{\substack{W \subseteq \mathbb{R}^2 \\ \text{всички крит.} \\ \text{области} \\ \alpha = \mathbb{P}(X \in W | H_0)}} \mathbb{P}(\vec{X} \notin W | H_1).$$

Постановка. X е случайна величина; $F_X(x, \theta)$ е разпределението на X , което зависи от накакъв параметър θ , но допускаме, че $f_X(x, \theta)$ е плътността на X (т.е. допускаме, че $\frac{\partial x}{\partial} F_X(x, \theta)$ съществува). Въвеждаме

$$f_{\vec{X}}(x, \theta) = \underbrace{L(X, \theta)}_{\text{ф-я на правдоподобие}} = \prod_{j=1}^n f_X(x_j, \theta), \text{ където } x \in \mathbb{R}^n \text{ и } x = (x_1, \dots, x_n).$$

Тогава е верен следния резултат:

Лема (Нейман-Пиърсън). Нека X удовлетворява горните условия от постановката и тестваме следната хипотеза:

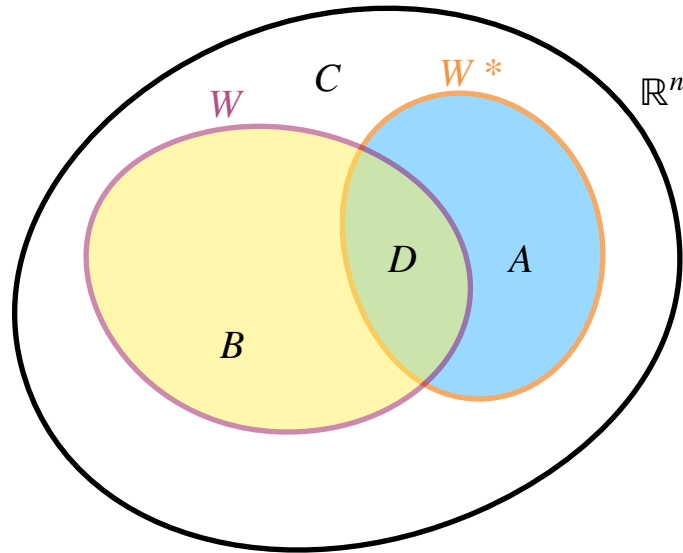
$$H_0 : \theta = \theta_0 \text{ срещу } H_1 : \theta = \theta_1.$$

Ако $L_0(x) = L(x, \theta_0)$ и $L_1(x) = L(x, \theta_1)$ и

$\exists k \geq 0 : W^* \subseteq \{x \in \mathbb{R}^n : L_1(x) \geq kL_0(x)\}$, $\overline{W^*} \subseteq \{X \in \mathbb{R}^2 : L_1(x) \leq kL_0(x)\}$ и $\alpha = \mathbb{P}(X \in W^* | H_0)$ е зададена, то W^* е ОКО (оптимална критична област).

Доказателство:

$$\underbrace{\alpha = \mathbb{P}(\vec{X} \in W^* | H_0) = \mathbb{P}(\vec{X} \in W | H_0)}_{\text{имаме}} \Rightarrow \underbrace{\mathbb{P}(\vec{X} \notin W^* | H_1) \leq \mathbb{P}(\vec{X} \notin W | H_1)}_{\text{искаме да докажем}}$$



$$\begin{aligned} \mathbb{P}(\vec{X} \notin W | H_1) &\stackrel{\theta=\theta_1}{=} \int_{\overline{W}} L_1(x) dx = \int_A L_1(x) dx + \int_C L_1(x) dx + \underbrace{\int_B L_1(x) dx - \int_B L_1(x) dx}_{=0} \\ &= \int_{\overline{W}^*} L_1(x) dx + \int_A L_1(x) dx - \int_B L_1(x) dx = \mathbb{P}(\vec{X} \in W^* | H_1) + \underbrace{\int_A L_1(x) dx - \int_B L_1(x) dx}_{\stackrel{?}{\geq 0}} \geq \end{aligned}$$

$\geq \mathbb{P}(\vec{X} \notin W^* | H_1)$, което искахме да докажем.

Т.е. всичко се свежда до това да проверим, че $\int_A L_1(x) dx - \int_B L_1(x) dx \geq 0$, но

$$\int_A L_1(x) dx - \int_B L_1(x) dx \geq \underbrace{k \int_A L_0(x) dx - k \int_B L_0(x) dx}_{\stackrel{?}{\geq 0}}$$

$$\alpha = \int_{W^*} L_0(x) dx = \int_A L_0(x) dx + \int_D L_0(x) dx = \int_W L_0(x) dx = \int_B L_0(x) dx + \int_D L_0(x) dx$$

$$\Rightarrow \int_A L_0(x) dx = \int_B L_0(x) dx.$$