СЕМ, лекция 14 (2021-01-14)

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)

$\overrightarrow{X}(X_1, \dots, X_n)$ - 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 $\underbrace{\hat{\theta} = \hat{\theta}(\overrightarrow{X})}_{\text{точкова оценка}}$, която да я вземем близо до θ, така че да определи това 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 нарачем *доверителен*.

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

Цел: Ще търсим две числа $I_1 = I_1(\overrightarrow{X}) < I_2 = I_2(\overrightarrow{X})$ такаива, 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 $T = T(\overrightarrow{X}, \theta)$ е централна статистика, ако:

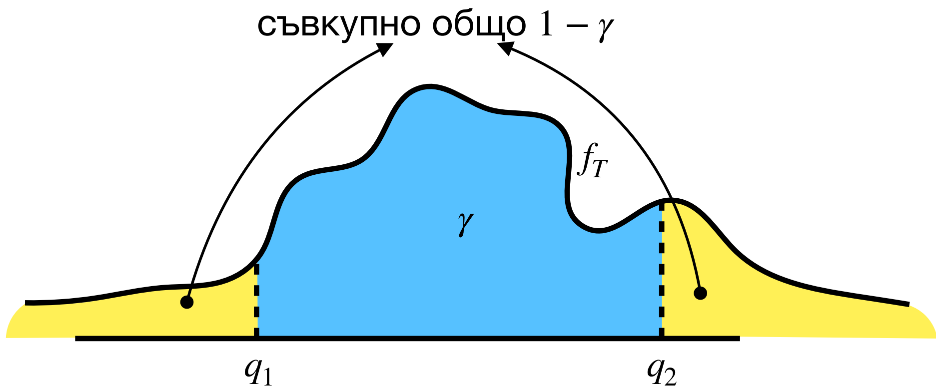
1) T е монотонна по θ
2) $\mathbb{P}(T < x) = F_T(x)$ не зависи от θ (T е функция на θ, но разпределението и не зависи от θ)

$\oplus$ Имаме $\overrightarrow{X}$ (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

$\gamma = \mathbb{P}(q_1 < T < q_2) \overset{\text{при фиксиран вектор на наблюдения } \overrightarrow{X}}{=} \mathbb{P}\left(T^{-1}(q_1) < \theta < T^{-1}(q_2)\right)$, защото сме допускали, че T расте по θ. Ако T намаляваше по θ, щяхме да имаме $\gamma = \mathbb{P}\left(T^{-1}(q_2) < \theta < T^{-1}(q_1)\right)$. Следователно интервала ще е $I_1 = T^{-1}(q_1)$, $I_2 = T^{-1}(q_2)$.

f_T - плътността на централната статистика (ЦС)

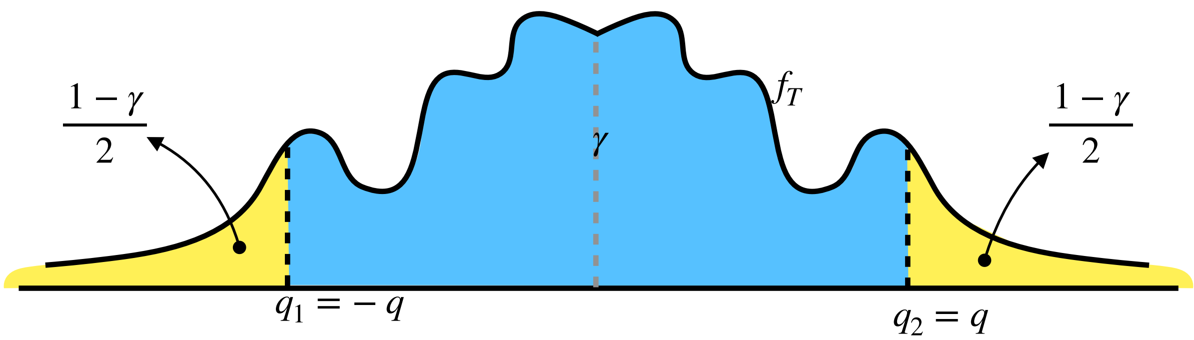


Има много начини, по които може да изберем q_1 и q_2, така че вероятността между тях да е γ.

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

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

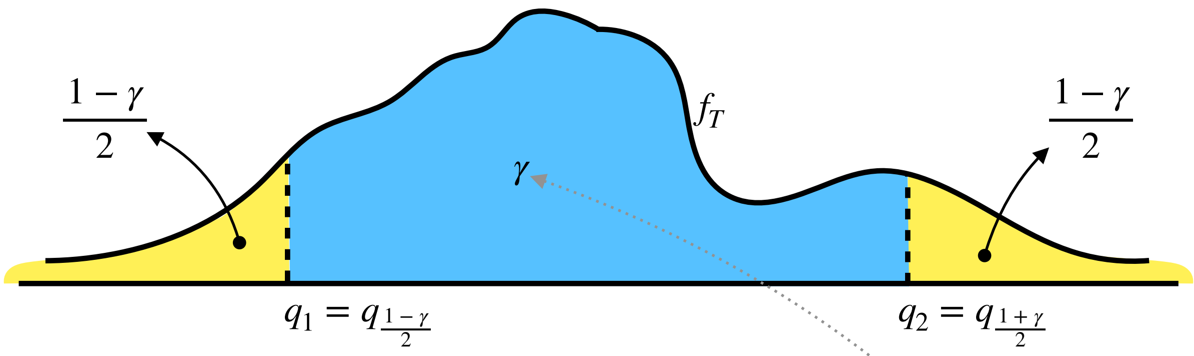
$\oplus$ Ако T е симетрична случайна величина,



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

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

<u>По-общо</u>: Ако имаме някакво несиметрично разпределение на T:



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

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.

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

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

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). Искаме да видим как може да оценим μ и да намерим за него доверителен интервал.

Ние знаем, че $\hat{\mu} = \overline{X}_n = \dfrac{1}{n}\sum_{j=1}^{n} X_j$. Тогава $T(\overrightarrow{X}, \mu) = \dfrac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \in \mathcal{N}(0,1)$, σ е известно число. Това е така, защото

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

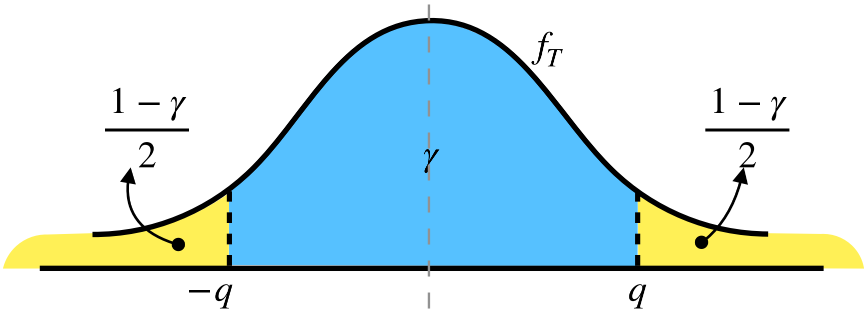
[По-подробно обяснение на свойството линейност на нормалното разпределение:

$$\overline{X_n} = \frac{\sum_{j=1}^{n} X_j}{n} \in \mathcal{N}(x, y^2).\ \mathbb{E}\overline{X_n} = \frac{n\mu}{n} = x;\ \mathrm{Var}(\overline{X_n}) = \frac{1}{n^2}\sum_{j=1}^{n}\sigma^2 = \frac{\sigma^2}{n} = y^2 \Big]$$

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

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.

$$\gamma = \mathbb{P}(-q < T < q) = \mathbb{P}\left(-q_{\frac{1}{2}+\frac{\gamma}{2}} < \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < q_{\frac{1}{2}+\frac{\gamma}{2}}\right) \Rightarrow$$

$$\Rightarrow \mathbb{P}\left(\mu \in (\underbrace{\overline{X_n} - \frac{\sigma}{\sqrt{n}} \times q_{\frac{1}{2}+\frac{\gamma}{2}}}_{I_1}, \underbrace{\overline{X_n} + \frac{\sigma}{\sqrt{n}} \times q_{\frac{1}{2}+\frac{\gamma}{2}}}_{I_2})\right) = \gamma \Rightarrow$$

$$\Rightarrow I_1 = \overline{X_n} - \frac{\sigma}{\sqrt{n}} \times q_{\frac{1}{2}+\frac{\gamma}{2}};\ I_2 = \overline{X_n} + \frac{\sigma}{\sqrt{n}} \times q_{\frac{1}{2}+\frac{\gamma}{2}}.$$

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

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

$\underbrace{\hat{\mu} = \overline{X_n}}_{\text{независимо дали знаем или не } \sigma}$ и оценката за дисперсията е $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{j=1}^{n}(X_j - \overline{X_n})^2$ и е неизместена.

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

<u>Твърдение</u>: Имаме, че $X \in \mathcal{N}(\mu, \sigma^2)$ и $\overrightarrow{X} = (X_1, \ldots, X_n)$ са наблюдения над X.
Тогава е вярно, че:

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

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

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

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

неизвестно

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

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

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

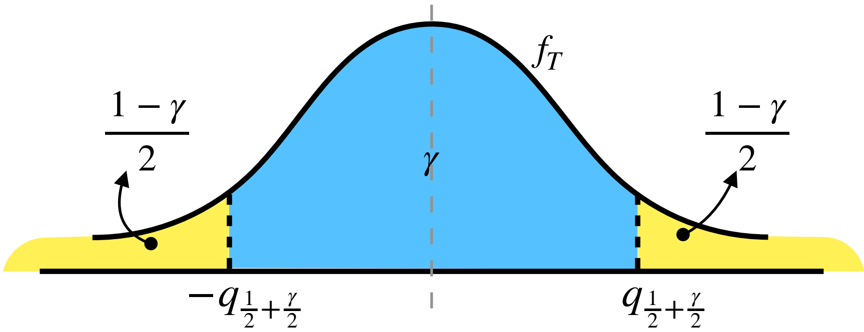
$\in \mathcal{X}^2(n+1)$

Заключения:

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

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

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



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

т.е. имаме, че I_1 и I_2 зависят от s и квантилите не са от нормалното разпределение а от $t-\text{Student's}$ разпределението.

Бележки: $T - \dfrac{\overline{X_n}-\mu}{\frac{s}{\sqrt{n}}} = \dfrac{Z}{\sqrt{\frac{Y}{n-1}}}, Z \perp\!\!\!\perp Y, Z \in \mathcal{N}(0,1), Y \in \mathcal{X}^2(n-1)$

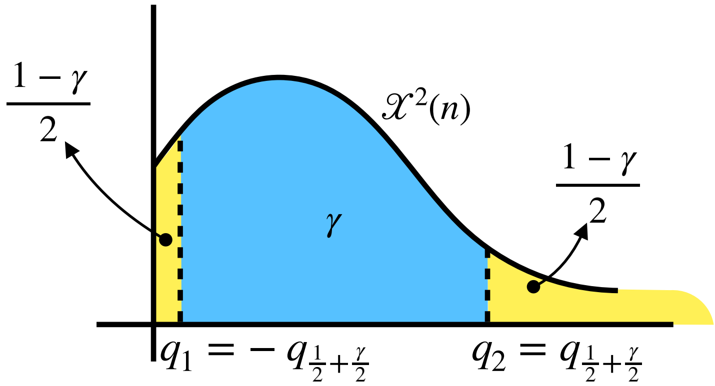
- $Y = \sum_{j=1}^{n-1} V_j$, където $(V_j)_{j=1}^{n-1}$ са независими и еднакво разпределени (i.i.d.) с $\mathcal{X}^2(1)$.
  От ЗГЧ: $\dfrac{Y}{n-1} \xrightarrow[n\to\infty]{\text{п.с.}} \mathbb{E}V_1 = 1$. Следователно за големи n : $T \approx \dfrac{Z}{1} \approx \mathcal{N}(0,1)$;
- Ако знаем дисперсията: $\dfrac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$ за произволни $X_1, \dots, X_n, \dots$
  $\mathbb{E}X_1 = \mu;\ \operatorname{Var}(X_1) = \sigma^2$.

$\oplus$ $\quad X \in \mathcal{N}(\mu, \sigma^2)$, знаем μ, $\hat{\sigma}^2 = \dfrac{1}{n}\sum_{j=1}^{n}(X_j-\mu)^2$. 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: $\dfrac{n\hat{\sigma}^2}{\sigma^2} = \sum_{j=1}^{n}\underbrace{\left(\dfrac{X_j-\mu}{\sigma}\right)}_{Z_j^2} = \sum_{j=1}^{n} Z_j^2$, 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

$T = \dfrac{n\hat{\sigma}^2}{\sigma^2}$ е централна по дефиниция (разпределението и е $\mathscr{X}^2(n)$ - не зависи от σ^2 и тя е монотонно намаляваща по σ).
За квантилите:

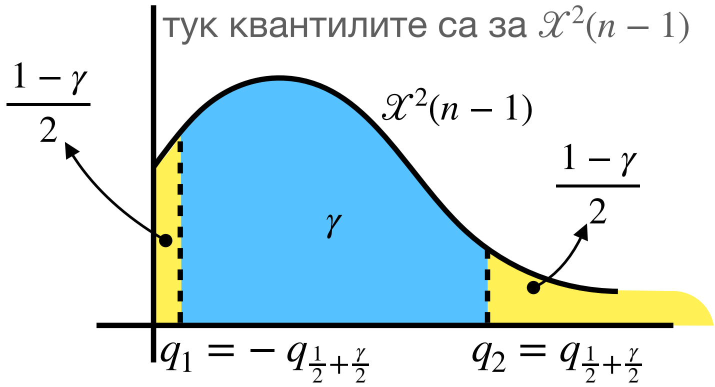


$$\gamma = \mathbb{P}(q_1 < T < q_2) = \mathbb{P}\left(\frac{n\hat{\sigma}^2}{q_2} < \frac{n\hat{\sigma}^2}{q_1}\right) \Rightarrow \begin{cases} I_1 = \frac{n\hat{\sigma}^2}{q_2} \\ I_2 = \frac{n\hat{\sigma}^2}{q_1} \end{cases}$$

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

$(n-1)s^2 = \displaystyle\sum_{j=1}^{n} (X_j - \overline{X_n})^2$ 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: $\dfrac{(n-1)s^2}{\sigma^2} = \dfrac{\sum_{j=1}^{n} (X_j - \overline{X_n})^2}{\sigma^2}$, но знаем от твърдението, че $T = \dfrac{(n-1)s^2}{\sigma^2} \in \mathscr{X}^2(n-1)$ (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)



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

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

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 $\overrightarrow{X}$ попадне в W $(\overrightarrow{X} \in W)$, тогава отхвърляме H_0 (в полза на H_1), ако вектора от наблюдения $\overrightarrow{X}$ 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.

<u>Цел</u>: При наблюденията $\overrightarrow{X} \in \mathbb{R}^n$, търсим да конструираме $W \subseteq \mathbb{R}^n$: ако $\overrightarrow{X} \in W$, то отхвърляме H_0 и приемаме H_1, а ако $\overrightarrow{X} \in \overline{W}$, то приемаме H_0.

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

Грешки, които може да допуснем:

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

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

$\oplus$ H_0 : дадена ваксина е вредна ($\theta = \theta_0$)
H_1 : ваксината не е вредна ($\theta = \theta_1$)

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^-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

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

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 $\dfrac{\partial x}{\partial} F_X(x, \theta)$ съществува). Въвеждаме

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

Тогава е верен следния резултат:

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

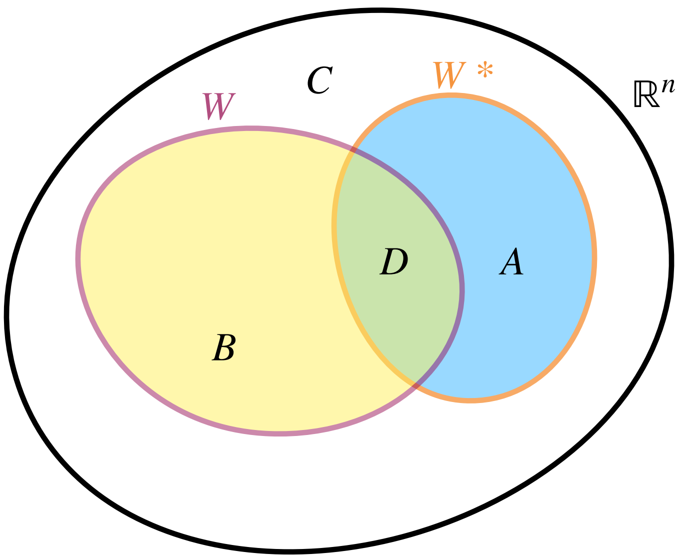
$H_0 : \theta = \theta_0$ срещу $H_1 : \theta = \theta_1$.

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:

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



$$\mathbb{P}(\overrightarrow{X} \notin W \,|\, H_1) \overset{\theta=\theta_1}{=} \int_{\overline{W}} L_1(x)\mathrm{d}\,x = \int_A L_1(x)\mathrm{d}\,x + \int_C L_1(x)\mathrm{d}\,x + \underbrace{\int_B L_1(x)\mathrm{d}\,x - \int_B L_1(x)\mathrm{d}\,x} =$$

$$= \int_{\overline{W\,*}} L_1(x)\mathrm{d}\,x + \int_A L_1(x)\mathrm{d}\,x - \int_B L_1(x)\mathrm{d}\,x = \mathbb{P}(\overrightarrow{X} \in W\,* \,|\, H_1) + \underbrace{\int_A L_1(x)\mathrm{d}\,x - \int_B L_1(x)\mathrm{dx}}_{\overset{?}{\geq} 0} \geq$$

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

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

$$\int_A L_1(d)\mathrm{d}\,x - \int_B L_1()x\,\mathrm{d}\,x \geq \underbrace{k\int_A L_0(x)\mathrm{d}\,x - k\int_B L_0(x)\mathrm{d}\,x}_{\overset{?}{=} 0}$$

$$\alpha = \int_{W*} L_0(x)\mathrm{d}\,x = \int_A L_0(x)\mathrm{d}\,x + \cancel{\int_D l_0(x)\mathrm{d}\,x} = \int_W L_0(x)\mathrm{d}\,x = \int_B L_0(x)\mathrm{d}\,x + \cancel{\int_D L_0(x)\mathrm{d}\,x}$$

$$\Rightarrow \int_A L_0(x)\mathrm{d}\,x = \int_B L_0(x)\mathrm{d}\,x.$$

github.com/andy489