СЕМ, лекция 15
(2021-01-21)

Припомняне.

- Имаме случайна величина X, която искаме „да разберем“, т.е. да извлечем някаква информация за нея, и чиято функция на разпределение зависи от някакъв параметър θ (едномерен);
- $\overrightarrow{X}$ са някакви наблюдения дадени като n-мерен вектор;
- α (алфа) е предварително зададена грешка от I$^{\text{-ви}}$ род;
- Тестваме две прости хипотези:
  $H_0 : \theta = \theta_0$ (базова)
  $H_1 : \theta = \theta_1$ (алтернативна)

Означаваме: $L_0(x) = L(x; \theta_0)$ (функция на правдоподобие за $\theta = \theta_0$) и $L_1(x) = L(x; \theta_1)$ (функция на правдоподобие за $\theta = \theta_1$).

Търсим $W^* \subseteq \mathbb{R}^n$ (област на $\mathbb{R}^n$), така, че когато нашия n-мерен вектор от наблюдения попадне в него, ние да отхвърляме нулевата хипотеза и да приемаме алтернативната. $\alpha = \mathbb{P}(\overrightarrow{X} \in W^* \,|\, H_0)$ е грешка от първи род, т.е. да попаднем в W^* и да отхвърлим нулевата хипотеза, но тя да е била вярна. Тази грешка е презададена (предефинирана) и се контролира от изследователя. Целта на оптималната критична област W^* е да намери тази област, която минимизира грешката от първи род.

$$\beta = \min_{\alpha = \mathbb{P}(\overrightarrow{X} \in W | H_0)} \mathbb{P}(\overrightarrow{X} \notin W \,|\, H_1).$$

Лемата на Нейман-Пиърсън е „добра“, защото ни характеризира даден критетии, по който да определим дали една област е оптимална критична област и по тази лема знаем, че W^* е о.к.о. (оптимална критична област), ако съществува някаква константа K (K може да зависи от n и от θ, но не може да зависи от x), за която

$$W^* \in \{x \in \mathbb{R}^n : L_1(x) > K \times L_0(x)\}$$
$$W^{*c} \in \{x \in \mathbb{R}^n : L_1(x) \leq K \times L_0(x)\}$$

и ако знаем, че е изпълнено равенството: $\alpha = \mathbb{P}(\overrightarrow{X} \in W^* \,|\, H_0)$, то W^* е о.к.о.

$\oplus$ $X \in \mathcal{N}(\mu, \sigma^2)$, където σ^2 е известно и искаме да построим оптимална критична област за тестване на хипотезата на μ (за намиране на средното, знаейки каква е дисперсията)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_1$$

Допускаме за улеснение, че $\mu_1 > \mu_0$. При зададено α.

$$L_0(x) = (\sqrt{2\pi}\sigma)^{-n} e^{-\frac{\sum_{j=1}^n (x_j - \mu_0)^2}{2\sigma^2}}, L_1(x) = (\sqrt{2\pi}\sigma)^{-n} e^{-\frac{\sum_{j=1}^n (x_j - \mu_1)^2}{2\sigma^2}}.$$

От лемата на Нейман-Пиърсън знаем, че оптималните критични области се намират лесно с неравенства от вида:

$$\{X \in \mathbb{R}^n : L_1(x) \geq K \times L_0(x)\} = \left\{ X \in \mathbb{R}^n : \frac{-\sum_{j=1}^n (x_j - \mu_1)^2}{2\sigma^2} \geq \ln K - \frac{\sum_{j=1}^n (x_j - \mu_0)^2}{2\sigma^2} \right\} =$$

$$= \left\{ X \in \mathbb{R}^n : -\cancel{\frac{\sum_{j=1}^n x_j^2}{2\sigma^2}} + \frac{1}{\sigma^2}\mu_1 \sum_{j=1}^n x_j - \underbrace{\frac{n\mu_1^2}{2\sigma^2}}_{\text{не зависи от наблюденията } \overrightarrow{X}} \geq \right.$$

$$\left. \geq -\cancel{\frac{\sum_{j=1}^n x_j^2}{2\sigma^2}} + \frac{1}{\sigma^2}\mu_0 \sum_{j=1}^n x_j - \underbrace{\frac{n\mu_0^2}{2\sigma^2}}_{\text{не зависи от наблюденията } \overrightarrow{X}} \right\} =$$

$$= \left\{ X \in \mathbb{R}^n : \frac{1}{\sigma^2} \underbrace{(\mu_1 - \mu_0)}_{\substack{\mu_1 > \mu_0 \\ \text{по допускане}}} \sum_{j=1}^n x_j \geq K_1 \right\} = \text{, където } K_1 = \ln K - \frac{n\mu_0^2}{2\sigma^2} + \frac{n\mu_1^2}{2\sigma^2}$$

$$= \left\{ X \in \mathbb{R}^n : \sum_{j=1}^n x_j \geq K_2 \right\} = \text{, където } K_2 = \frac{K_1\sigma^2}{\mu_1 - \mu_0}$$

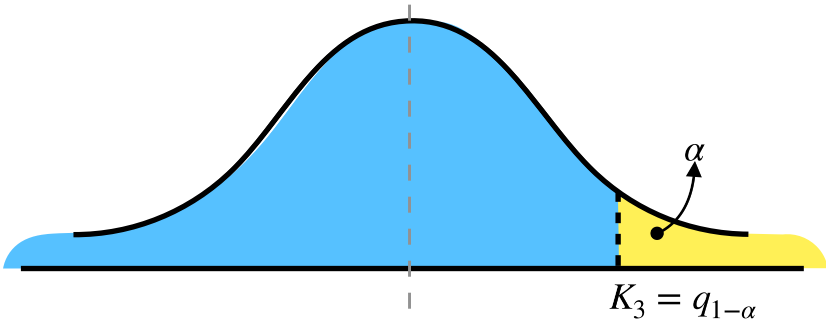
$$= \left\{ X \in \mathbb{R}^n : \overset{=\overline{X}}{\frac{\sum_{j=1}^m x_j}{n}} \geq \frac{K_2}{n} \right\} = \left\{ X \in \mathbb{R}^n : \overline{X} \geq \frac{K_2}{n} \right\} =$$

$$= \left\{ X \in \mathbb{R}^n : \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{K_2}{\sigma\sqrt{n}} = K_3 \right\}$$

$$\alpha = \mathbb{P}(\overrightarrow{X} \in W \,|\, H_0) = \mathbb{P}\big(\underbrace{L_1(\overrightarrow{X}) \geq K_0 L_0(\overrightarrow{X})}\,|\, H_0\big) = \mathbb{P}\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq K_3 \,|\, H_0 \right) \quad ,$$

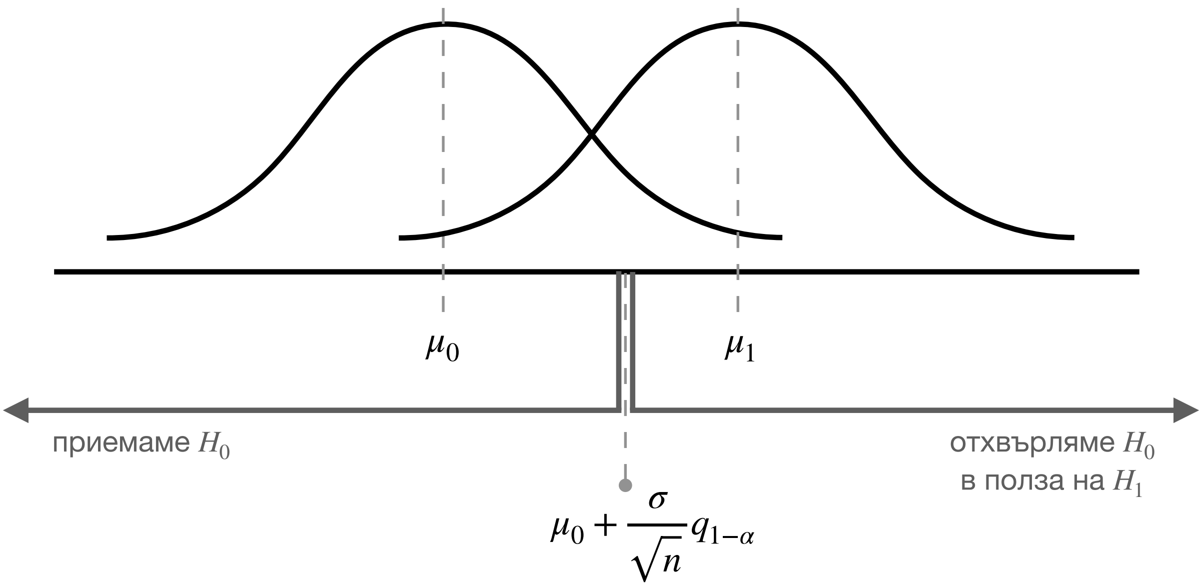
където $\overline{X} = \dfrac{\sum_{j=1}^n x_j}{n}$.

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \in \mathcal{N}(0,1) = \mathbb{P}(Z \geq K_3)$$



$$\Rightarrow K_3 = q_{1-\alpha}$$

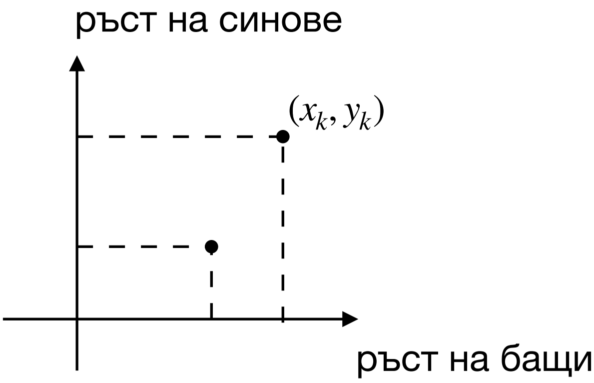
о.к.о.: $\left\{ \overline{X} \geq \mu_0 + q_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} \right\}$



Линейна регресия

(Галтон)

Нека a е средния ръст на мъжете.

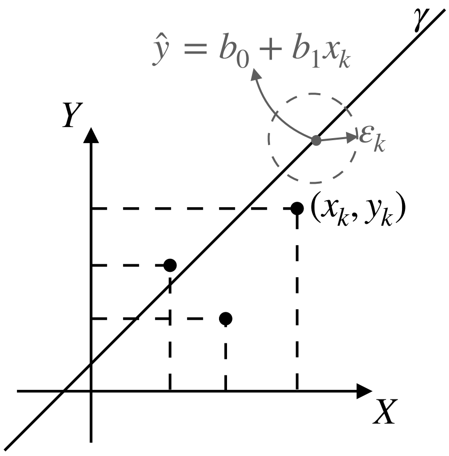


$y_{\text{син}} = a + \beta(x_{\text{баща}} - a)$.

Галтон е забелязал, че по неговите данни, коефициента бета е $\beta = 0.6$. Тоест, ако бащата е 10 см. над средния ръст, то сина му ще е с 6 см. над средния ръст. Тази по-слаба зависимост е влязла в теорията като регрес (завръщане) към средното. Синовете на високите бащи не са чак толкова високи в средно както бащите им, а са на около половината от отклонението на бащата над средния ръст за мъжете.

Модел: $Y = \beta_0 + \beta_1 X + \varepsilon$ грешка

отклик предиктор



Допускаме че в множеството от точки $(x_1, x_2, \ldots, x_n)$ и $(y_1, y_2, \ldots, y_n)$ има някакъв линеен модел. Т.е. предполагаме, че има линеен модел $y_k = b_0 + b_1 x_k + \varepsilon_k$. Тоест имаме някаква права γ (от чертежа). Искаме да си построим линеен модел, а не някакъв друг, за да не рискуваме да интерполираме, тъй като интерполацията няма добра статистическа стойност. Т.е. не е добре да обхванем всички данни с много сложна крива и в момента, в който добавим данни – кривата ни да е твърде динамична и да няма никаква прогнозна сила. Интерполацията може да мине през n точки, но при добавянето на $n+1$-вата точка – точността на кривата да рухне.

$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} x_k$ и $\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} y_k$.

Търсим: $\min\limits_{b_0,\, b_1} \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 = \min\limits_{b_0,\, b_1} \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k)^2$.

(С квадратични грешки се смята по-лесно, а освен това имат и статистическо значение. Въпреки това тук може да имаме най-разнообразни метрики, които искаме да оптимизираме (например абсолютната стойност или максималното отклонение измежду всички възможни отклонения и т.н.))

Искаме да минимизираме функцията по-горе по две променливи. За целта ще си веземем производната по b_0 и тя трябва да бъде нула и аналогично за производната по b_1:

$$0 = \frac{\partial}{\partial b_0} \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k)^2 = -2 \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k) \Rightarrow$$

$$\Rightarrow \sum_{k=1}^{n} y_k - n b_0 - b_1 \sum_{k=1}^{n} x_k = 0 \Rightarrow n\overline{Y} - n b_0 - n b_1 \overline{X} = 0 \Rightarrow \overline{Y} = b_0 + b_1 \overline{X}$$

$$0 = \frac{\partial}{\partial b_1} \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k)^2 = -2 \sum_{k=1}^{n} (b_0 + b_1 x_k - y_k) x_k \Rightarrow$$

$$\Rightarrow n\overline{X}(\overline{Y} - b_1 \overline{X}) + b_1 \sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k y_k$$

$$\begin{cases} b_0 = \overline{Y} - b_1 \overline{X} \\ b_1 = \dfrac{\sum_{k=1}^{n} x_k y_k - n\overline{Y}\overline{X}}{\sum_{k=1}^{n} x_k^2 - n(\overline{X})^2} = \dfrac{\sum_{k=1}^{n} (y_k - \overline{Y})(x_k - \overline{X})}{\sum_{k=1}^{n} (x_k - \overline{X})^2} = \dfrac{\sum_{k=1}^{n} (x_k - \overline{X}) y_k}{\sum_{k=1}^{n} (x_k - \overline{X})^2} \end{cases}$$

$$A = \sum_{k=1}^{n} (X_k - \overline{X})^2$$

Допускаме, че Y_k като отговор на X_k е случайна величина, в смисъл, че съществуват неизвестни коефициенти β_0, β_1, които при зададено X_k дават следната линейна зависимост, където ε_k е случайна грешка.
Допускаме, че $Y_k = \beta_0 + \beta_1 X_k + \varepsilon_k$, където $k = \overline{1,n}$.

Правим и следните допускания за епсилон грешките:
$(\varepsilon_i)_{i=1}^{n}$ са независими еднакво разпределени случайни величини, като $\varepsilon_i \in \mathcal{N}(0, \sigma^2)$

хомоскедастичност

Т.е. грешките са нормално разпределени и независими една от друга. Т.е. нямаме системна грешка. Хомоскедастичността е малко по-тежко допускане, но тя придава простота на модела.

$$Y_k \in \mathcal{N}(\beta_0 + \beta_1 X_k, \sigma^2)$$

$$\begin{cases} \hat{\beta}_0 = b_0 = \overline{Y} - b_1 \overline{X} \\ \hat{\beta}_1 = b_1 = \underbrace{\dfrac{\sum_{k=1}^{n} X_k Y_k - n\overline{X}\overline{Y}}{\sum_{k=1}^{n} X_k^2 - n(\overline{X})^2}}_{(1)} = \underbrace{\dfrac{\sum_{k=1}^{n} (Y_k - Y)(X_k - X)}{\sum_{k=1}^{n} (X_k - \overline{X})^2}}_{(2)} = \underbrace{\dfrac{\sum_{k=1}^{n} (X_k - \overline{X}) Y_k}{\sum_{k=1}^{n} (X_k - \overline{X})^2}}_{(3)} \end{cases}$$

$A = \sum_{k=1}^{n} (X_k - \overline{X})^2$ си остава същото, тъй като е фиксирано число в знаменателя.

$$\mathbb{E}b_1 \overset{(3)}{=} \frac{1}{A}\sum_{k=1}^{n}(X_k - \overline{X})\mathbb{E}Y_k = \frac{1}{A}\sum_{k=1}^{n}(X_k - \overline{X})(\beta_0 - \beta_1 X_k) =$$

$$= \frac{\beta_0}{A}\underbrace{\cancel{\sum_{k=1}^{n}(X_k - \overline{X})}}_{=0} + \frac{\beta_1}{A}\underbrace{\sum_{k=1}^{n}(X_k - \overline{X})X_k}_{=A} = \beta_1 .$$

Оказва се, че очакването на b_1 е равно на β_1, което ни казва, че b_1 е неизместена оценка на β_1.

$$\mathbb{E}b_0 = \mathbb{E}\overline{Y} - \overline{X}\mathbb{E}b_1 = \frac{1}{n}\sum_{k=1}^{n}Y_k - \beta_1\overline{X} =$$

$$= \frac{1}{n}\sum_{k=1}^{n}(\beta_0 + \cancel{\beta_1 X_k}) - \cancel{\beta_1\overline{X}} = \beta_0 .$$

И b_0 и b_1 са неизместени оценки на неизвестните парамвтри β_0 и β_1. По този начин знаем, че нямаме систематична грешка, когато правим тези оценки.

$$\mathbb{D}b_1 \overset{(3)}{=} \mathbb{D}\frac{\sum_{k=1}^{n}(X_1 - \overline{X})Y_k}{A} = \frac{1}{A^2}\sum_{k=1}^{n}(X_n - \overline{X})^2\underbrace{\mathbb{D}Y_k}_{\sigma^2} =$$

$$= \sigma^2\frac{\cancel{\sum_{k=1}^{n}(X_n - \overline{X})^2}}{A^{\cancel{2}}} = \frac{\sigma^2}{A} .$$

$$\Rightarrow b_1 = \underset{\text{оценка}}{\hat{\beta}_1} \in \mathcal{N}\left(\beta_1, \frac{\sigma^2}{A}\right)$$

Това означава, че вече може да тестваме хипотези за b_1.

За дисперсията на b_0 по същата логика може да докажем, че:

$$\mathbb{D}b_0 = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{A}\right) \Rightarrow b_0 = \underset{\text{оценка}}{\hat{\beta}} \in \mathcal{N}\left(\beta_0, \sigma\left(\frac{1}{n} + \frac{\overline{X}^2}{A}\right)\right).$$

Двете дисперсии клонят към нула.

Оценка на σ^2 (ако не го знаем априорно). $Y_k \in \mathcal{N}(\beta_0 + \beta_1 X_k, \sigma^2)$. Проблема е, че не знаем β_0 и β_1, тъй като, ако допуснем, че ги знаем, щяхме да имаме $\dfrac{Y_k - \beta_0 - \beta_1 X_k}{\sigma} \in \mathcal{N}(0,1)$ и тогава

$$\frac{\sum_{k=1}^{n}(Y_k - \beta_0 - \beta_1 X_k)^2}{\sigma^2} \in \mathcal{X}^2(n).$$

Но, ако са ни верни допусканията за модела, тогава:

$\frac{\sum_{k=1}^{n}(Y_k - b_0 - b_1 X_k)^2}{\sigma^2} \in \mathcal{X}^2(n-2)$ („изхабили" (използвали) сме две степени на свобода (две данни), за да оценим b_0 и b_1)

$$\mathbb{E}\frac{\sum_{k=1}^{n}(Y_k - b_0 - b_1 X_k)^2}{\sigma^2} = n - 2$$

$\hat{\sigma}^2 = \frac{\sum_{k=1}^{n}(Y_k - b_0 - b_1 X_k)^2}{n-2}$, т.е. $\sigma^2 = \mathbb{E}\hat{\sigma}^2$.

Оттук нататък ние може да тестваме хипотези. Може да си конструираме множество хипотези от следния вид:

$$H_0 : \beta_1 = \tilde{\beta}$$
$$H_1 : \beta_1 = \tilde{\tilde{\beta}}$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$\frac{b_1 - \tilde{\beta}}{\sqrt{\sigma^2/A}} \in \mathcal{N}(0,1)$ при H_0.