15.12.2025

lundi, 15 décembre 2025 08:10

Rappel: mardi 16.12 : séance réponse aux questions.

Rappel sur les 2-samples t-test :

$X_1, \dots, X_n$, $Y_1, \dots, Y_n$ deux échantillons

$$T_n(X_1, \dots, X_n, Y_1, \dots, Y_n) = \frac{\sqrt{n}\,(\bar{X}_n - \bar{Y}_n)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}}$$

Hypothèses: X_i, Y_i sont gaussiens.

de même variance

↳ motivation F-test

F-tests :

Déf: (lois de Fisher) Soient $k_1, k_2 \geq 1$ deux entiers.

Une variable aléatoire X suit une loi de Fisher à (k_1, k_2)-degrés de liberté, $X \sim \text{Fisher}(k_1, k_2)$, si c'est une variable continue avec densité :

$$f_X(x) = \mathbb{1}_{[0,\infty)}(x) \; \frac{1}{\quad} \; \frac{1}{x\,(\quad)} \left(\frac{k_1 x}{\quad}\right)^{\frac{k_1}{2}} \left(1 - \frac{k_1 x}{\quad}\right)^{\frac{k_2}{2}}$$

$$f_X(x) = \mathbb{1}_{[0,+\infty)}(x) \; \frac{1}{B(k_1/2, k_2/2)} \; \frac{1}{x} \left(\frac{k_1 x}{k_1 x + k_2}\right)^{\frac{k_1}{2}} \left(1 - \frac{k_1 x}{k_1 x + k_2}\right)^{\frac{k_2}{2}}$$

où $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$.

Intérêt : Si $X \sim \chi^2_{k_1}$, $Y \sim \chi^2_{k_2}$ sont indép., alors

$$\frac{X/k_1}{Y/k_2} \sim \text{Fisher}(k_1, k_2).$$

Vu en exo :

$$E[X] = \frac{k_2}{k_2 - 2} \quad \text{si } X \sim \text{Fisher}(k_1, k_2)$$

$$\text{Var}(X) = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2-2)^2(k_2-4)} \overset{k_1 = k_2 = n}{\sim} \frac{1}{n}$$

2 tests dans le cadre ANOVA

test 1 : F-test pour l'égalité de variance

Rappel : $X_1, \dots, X_n$ un échantillon de $\mathcal{N}(\mu, \sigma^2)$

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}.$$

$$\hookrightarrow \frac{1}{(n-1)\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \sim \frac{\chi^2_{n-1}}{n-1}$$

Cadre : 2 échantillons $X_1, \dots, X_n$ et $Y_1, \dots, Y_m$ de lois gaussiennes : $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ et $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.

On suppose $(X_i)_i$ et $(Y_i)_i$ sont indép.

But : tester H_0 : "$\mathrm{Var}(X_1) = \mathrm{Var}(Y_1)$" = "$\sigma_X^2 = \sigma_Y^2$"

$\Theta = \mathbb{R} \times \mathbb{R} \times (0, \infty) \times (0, +\infty)$, $\theta \in \Theta$

$\theta = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2)$

$\Theta_0 = \{(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2) \in \Theta : \sigma_X^2 = \sigma_Y^2\}$.

↳ on va utiliser la statistique de test :

$$T(X_1, \dots X_n, Y_1, \dots, Y_m) = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2}{\frac{1}{m-1} \sum_{j=1}^{m} (Y_i - \overline{Y}_m)^2}$$

Si H_0 est vraie, $\sigma_Y^2 = \sigma_X^2$

$$T(X_1, \dots X_n, Y_1, \dots, Y_m) = \frac{\frac{1}{(n-1)\sigma_X^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2}{\frac{1}{(m-1)\sigma_Y^2} \sum_{i=1}^{m} (Y_i - \overline{Y}_m)^2} \cdot \frac{\sigma_X^2}{\sigma_Y^2}$$

Sous H_0, $\frac{\sigma_X^2}{\sigma_Y^2} = 1$,

$\sim$ Fisher $(n-1, m-1)$

Fisher $(n-1, m-1)$

Si H_1 est vraie : $\sigma_X^2 \neq \sigma_Y^2$

$$T(X_1, \dots, X_n, Y_1, \dots, Y_m) \sim \frac{\sigma_X^2}{\sigma_Y^2} \text{ Fisher}(n-1, m-1)$$

Si $n, m \gg 1$, Fish$(n-1, m-1)$ se concentre autour de 1

$\rightsquigarrow$ Sous H_0 : $T \in [1-\varepsilon, 1+\varepsilon]$ avec très grande proba

Sous H_1 : $|T-1| > 0$ confiance sur n, m avec grande proba.

$\rightsquigarrow$ On va utiliser une région de rejet de la forme

$$\{ T(X_1, \dots X_n, Y_1, \dots, Y_m) \notin [1-\delta_\alpha, 1+\delta'_\alpha] \}$$

avec $\delta_\alpha, \delta'_\alpha$ t.q.

- $\mathbb{P}(\text{Fisher}(n-1, m-1) \leq 1-\delta_\alpha) = \frac{\alpha}{2}$
- $\mathbb{P}(\text{Fisher}(n-1, m-1) \geq 1+\delta'_\alpha) = \frac{\alpha}{2}$

Sous H_0 :

$$\sup_{\theta \in \Theta_0} \mathbb{P}(T(X_1, \dots X_n, Y_1, \dots, Y_m) \notin [1-\delta_\alpha, 1+\delta'_\alpha]) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

$\rightsquigarrow$ test de niveau de risque α.

Commande utile : si vous utilisez R :

Trouver z t.q. $\mathbb{P}(\text{Fisher}(n,m) \leq z) = \alpha$

↪ $qf(\alpha, n, m)$

2ème test : "one way ANOVA"

But : tester si l'appartenance à un certain groupe modifie la valeur d'un paramètre.

Idée : on a une population Pop (ensemble fini, très grand). On la divise en groupes

$$Pop = G_1 \cup \dots \cup G_k \quad , \quad k \geq 2$$

$$G_i \cap G_j = \emptyset \quad \text{ssi} \quad i \neq j.$$

Par exp! : $Pop = \{\text{étudiants EPFL}\}$

$G_i = \{$ étudiants qui passe entre $i-1$ et i heures / jour sur Tik Tok $\}$.

On s'intéresse à une certaine quantité sur Pop.

exp! : - nombre heures sommeil / nuit

exp! : - nombre heures sommeil / nuit
- la note moyenne.

But : déterminer si l'appartenance à différents groupes a cette influence sur la valeur de la quantité d'intérêt.

Formellement : On suppose que l'on a

Pop un ens., $Pop = G_1 \sqcup \dots \sqcup G_\ell$,

On a X_i, $i \in Pop$ une famille de v.a. i.i.d.

On suppose : $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$

$\mu_i \in \mathbb{R}$, $i \in Pop$, $\sigma^2 \in (0, +\infty)$ est la même pour tous les i.i.d.

$\Theta = \mathbb{R}^{Pop} \times (0, +\infty)$, $\theta \in \Theta$, $\theta = \left((\mu_i, i \in Pop), \sigma^2\right)$

On va tester l'hypothèse H_0 " les moyennes de chaque groupe sont égales à la moyenne générale de la population ".

$\Theta_0 = \left\{(\mu_i, i \in Pop, \sigma^2) : \frac{1}{|G_1|} \sum_{i \in G_1} \mu_i = \frac{1}{|G_2|} \sum_{i \in G_2} \mu_i = \dots\right.$

$$= \frac{1}{|Pop|} \sum_{i \in Pop} \mu_i \Big\}$$

On introduit :

les échantillons $X_{1,1}, \dots, X_{1,n_1}$ dans G_1

$X_{2,1}, \dots, X_{2,n_2}$ dans G_2

$\vdots$

$X_{k,1}, \dots, X_{k,n_k}$ dans G_k

$N = n_1 + n_2 + \dots + n_k$.

$\overline{X_{\ell,n_\ell}} = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} X_{\ell,i}$ moyenne empirique G_ℓ.

$\overline{X} = \frac{1}{N} \sum_{\ell=1}^{k} \sum_{i=1}^{n_\ell} X_{\ell,i}$ moyenne générale.

$$\text{In Grp Var} = \frac{1}{N-k} \sum_{\ell=1}^{k} \sum_{i=1}^{n_\ell} \left(X_{\ell,i} - \overline{X_{\ell,n_\ell}} \right)^2$$

$$\text{Bet Grp Var} = \frac{1}{k-1} \sum_{\ell=1}^{k} n_\ell \left(\overline{X_{\ell,n_\ell}} - \overline{X} \right)^2$$

Statistique de test : $T(-) = \dfrac{\text{Bet Grp Var}}{\text{In Grp Var}}$

Sous H_0, on a que toutes les moyennes sont identiques

$\rightsquigarrow$ $X_{\ell,i} - \overline{X_{\ell,n_\ell}}$ "$=$" (moralement) $\mathcal{N}(0, \sigma^2)$

et $\overline{X_{\ell,n_\ell}} - \overline{X}$ "$=$" (moral) $\mathcal{N}(0, \frac{\sigma^2}{n_\ell})$

et $\overline{X}_{\ell,n_\ell} - \overline{X}$ "$\approx$" (moral) $\mathcal{N}\left(0, \frac{\sigma^2}{n_\ell}\right)$

$\rightsquigarrow$ T est d'ordre 1.

Il se trouve que sous H_0,

$$T(-) \sim \text{Fisher}(k-1, N-k)$$

Sous H_1 : on a au moins un $\ell^* \in \{1, \dots, k\}$

tel que $\underbrace{\frac{1}{|G_{\ell^*}|} \sum_{i \in G_{\ell^*}} \mu_i}_{\overline{\mu}_{\ell^*}} \neq \underbrace{\frac{1}{|Pop|} \sum_{i \in Pop} \mu_i}_{\overline{\mu}}$

$$\text{Bet Grp Var} = \frac{1}{q-1} \sum_{\ell=1}^{q} n_\ell \left(\overline{X}_{\ell,n_\ell} - \overline{X}\right)^2$$

$$\geq \frac{1}{k-1} n_{\ell^*} \underbrace{\left(\overline{X}_{\ell^*,n_{\ell^*}} - \overline{X}\right)^2}$$

$$\xrightarrow[n_1, \dots, n_k \to \infty]{} \left(\overline{\mu}_{\ell^*} - \overline{\mu}\right)^2 > 0$$

$\hookrightarrow$ Sous H_1, T est la moitié d'un nombre >0 avec $n_\ell \approx \frac{N}{k} \xrightarrow{N \to \infty} \infty$.

Suggère la région de rejet

$$\left\{ \frac{\text{Bet Grp Var}}{\text{In Grp Var}} > \delta_\alpha \right\}$$

δ_α t.q. $\mathbb{P}\left(\text{Fisher}(k-1, N-k) > \delta_\alpha\right) = \alpha$.