Test de Student et Comparaison de Moyennes

1. Contexte Général :

- On s'intéresse à un phénomène que l'on étudie à l'aide de mesures. Ex : effet d'un médicament sur une charge virale.
- En général on étudie l'effet de facteurs sur une variable : Fact : médicament, Var : charge.
- On voudrait avoir une réponse qui soit généralisable à la pop. alors que les mesures sont faites uniq^t sur un ensemble restreint d'individus.
- Pose le Pb : - choix des individus : échantillonnage.
  - erreur commise lors de la généralisat°

		population Effet	Pas d'effet
échantillon	Effet	TP	FP
	Pas d'effet	FN	TN

- la stratégie des tests = tjs partir de l'hyp. "pas d'effet" = hypothèse nulle (raison^t par l'absurde).

→ on préfère dire qu'il n'y a pas d'effet plutôt que de se tromper. → Favorise plutôt H_0.

- Démarche de base : test d'hypothèse de comparaison de moyennes.

→ Comp. de populations dans un cadre limité.

ex : . dose du médicament ?
. facteurs qualit. quanti.

↳ Dans un 2^e^ tps, le modèle linéaire permettra de prendre en compte + appareils ds un cadre unifié.

②. Construction de la Stat. de Student.

- 2 pop A, B, Y^A, Y^B, $\overline{Y^A}$, $\overline{Y^B}$
- 3 infos : effectifs, moy, variance.
- la variance peut être connue ou non.

$$T = \frac{\overline{Y}_A - \overline{Y}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$ = quantité qui résume l'info et qui dépend des données.

↳ elle va servir à définir une règle de Décision quant au rejet de H_0.
= pour répondre à la question "y-a-t-il un effet ?"

↳ C'est un score.

- plus le score est élevé, plus on dira que la différence est significative → il faut un seuil
- Or on voit bien que l'erreur va dépendre du seuil : plus le seuil est petit, plus on aura FP.
  ↳ Comment choisir le seuil ?

③. L'utilité du modèle : Cas Gaussien à Variances Connues.

- le score de Student peut être utilisé indép. d'un trait^t statistique. Sa construct^° est naturelle et son interprétation intuitive.
- Si on modélise les données avec un modèle stat, alors on aura un cadre pour le choix du seuil.

Modèle Gaussien : - avec les hyp.
- loi des estimateurs.

sous H_0 : $\{\mu_A = \mu_B\}$ = l'hyp. concerne les param. du modèle.

$$T \underset{\mathcal{H}_0}{\sim} \mathcal{N}(0,1).$$

→ on sait faire les calculs sous $\mathcal{H}_0$.
→ on ne sait pas ce qu'il se passe sous $\mathcal{H}_1$.

- Règle de décision : On observe une valeur du score t_{obs} = une réalisation, et on la compare au Quantile de la loi normale.

Si $t_{obs} > u_{1-\alpha/2}$ alors on rejette $\mathcal{H}_0$.

↳ définit la région de rejet = $\mathcal{R}_\alpha$ = ensemble des valeurs de t irréalistes sous $\mathcal{H}_0$.

α = risque de 1ère espèce = 5%.

↳ nécessite la valeur du Quantile.

- Notion de P_v = $\mathbb{P}(|t_{obs}| \geq u_{1-\alpha/2})$

$= \mathbb{P}(|T| \geq t_{obs})$.

$(\mathbb{P}_v(t_{obs}) \leq \alpha) \Leftrightarrow (t_{obs} \in \mathcal{R}_\alpha)$

↓

reflète le t^x d'erreur que l'on est prêt à accepter. = niveau du test pour lequel H0 serait juste rejetée.

→ (α, β) liés : on fixe α et on max. selon π.

↳ une manière facile de comparer des tests

↳ selon leur π.

→ Comment max π ?

→ sous $\mathcal{H}_0$ $\quad \mathbb{E}(T) = 0, \quad \mathbb{V}(T) = \frac{n}{n-2}$

↳ si $n \nearrow$ on pourra mieux rejeter

→ Etape d'après : Calculer n pour π.

↳ suppose qu'on connaisse $\mathcal{H}_1$.

$\delta = \frac{\mu_A - \mu_B}{\sigma} \sqrt{R(R-1)}$ = diff normalisée.

↳ Reformuler la question :

- quelle π pour un n donné ?
- si on veut π, combien de n étant donné $\frac{\mu_A - \mu_B}{\sigma}$.

dans R : $\begin{cases} \text{power.test} \\ \text{t.test} \end{cases}$

Résumé sur la méthodologie.

- 4 étapes :
  - données $y_1 \dots y_n$
  - modèle définit la loi des Y, il dépend d'un paramètre θ.
  - hyp. à tester : hypothèse nulle qui porte sur les param.
  - règle de décision . une stat $T = f(Y_i)$
    - région de rejet.

④. Test de Student : modèle gaussien à variances égales.
↳ test sur les moyennes (construire le test).

⑤. Le prob. des données app.

<u>Hyp</u> : . les prop. A et B sont ⊥.
. si non vérifiée ?

→ on travaille sur les ≠.

⑥. Notion de puissance d'un test.

= π = proba que la procédure rejette H_0 quand elle est fausse.

Comparaison de populations (stat descriptive).

Modèle :

$$\begin{cases} \cdot\ x_{1A} \cdots x_{n_A A},\quad X_{1A} \cdots X_{n_A A},\ \text{iid}\ \mathcal{N}(\mu_A, \sigma^2) \\ \cdot\ x_{1B} \cdots x_{n_B B},\quad - \quad - \quad - \quad - \quad (\mu_B, \sigma^2) \\ \cdot\ \text{les 2 populat}^{\circ}\ \text{st}\ \perp\!\!\!\perp\ (\text{pas d'appariement}). \end{cases}$$

Hypothèse : $H_0 = \{\mu_A = \mu_B\}$.

→ Variances homogènes : $\cdot\ S^2 = \dfrac{1}{n_A + n_B - 2}\left((n_A - 1)s_A^2 + (n_B - 1)s_B^2\right)$

$$\cdot\ T = \frac{|\bar{X}_A - \bar{X}_B|}{S\sqrt{1/n_A + \frac{1}{n_B}}} \sim \mathcal{T}(n_A + n_B - 2)$$

→ Variances hétérogènes : $\cdot\ s_A^2 = \dfrac{1}{n_A - 1}\sum (X_{iA} - \bar{X}_A)^2$.

$$\cdot\ T = \frac{|\bar{X}_A - \bar{X}_B|}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \sim \mathcal{T}(df)$$

$$df = \frac{(w_A + w_B)^2}{\frac{w_A^2}{n_A - 1} + \frac{w_B^2}{n_B - 1}}, \quad w_A = \frac{s_A^2}{n_A}.$$

avec df approché par Satterswaite

Hypothèse préalable : $H_0 = \{\sigma_A^2 = \sigma_B^2\}$.

$$F = \frac{s_A^2 / n_A - 1}{s_B^2 / n_B - 1} \sim \mathcal{F}(n_A - 1, n_B - 1).$$

$$F' = \text{folded } F = \frac{\max(s_A^2, s_B^2)}{\min(s_A^2, s_B^2)}.$$

Rq : $IC(\sigma) = \left\{ \chi^2_{\alpha/2, n-1} \leq \dfrac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2, n-1} \right\}$.

Règle de décision :

(1). Si $\left\{ |t_{obs}| > |t_{1-\alpha/2, n-1}| \right\}$ $\Rightarrow$ Rejet avec risque α
$= \mathcal{R}_\alpha$ (région de rejet).

(2). Si $P_v = \mathbb{P}\left(|T| > |t_{obs}| \right) < \alpha$.

= proba d'obtenir les données si H_0 était vraie.

(1) $\Leftrightarrow$ (2).