Échantillons appariés

- On donne quelques exemples pour motiver les types de situation.
- Exemple 1. On est en présence de N couples de sujets homogènes. Le premier sujet du couple i reçoit le traitement A, disons, alors que le second sujet reçoit le traitement B. Exemples de sujets:
  1. Deux plantes de la même plante mère;
  2. Des jumeaux;
  3. Les deux yeux d'un même individu.
- Exemple 2. Mesures répétées. On est en présence de N sujets. Le sujet i est observé *avant* traitement, ainsi qu'*après* traitement.

Hypothèses

- Les deux exemples sont assez différents.
- Notons les N paires (X_i, Y_i), $i = 1, \ldots, N$.
- Dans le premier exemple, il peut être raisonnables de présumer que X_i et Y_i sont indépendantes.
- Dans le second exemple, elles sont dépendantes en général, car il s'agit du même sujet.

Introduction des différences

- On introduit les différences:

$$\Delta_i = Y_i - X_i, \; i = 1, \ldots, N;$$

- Si les N différences $\Delta_1, \ldots, \Delta_N$ proviennent d'une distribution F_Δ continue et strictement croissante, de médiane $Q_\Delta(1/2) = F_\Delta^{-1}(1/2)$, on aura alors:

$$P(\Delta = Q_\Delta(1/2)) = 0.$$

- On pose $E(\Delta) = \theta$.

Le test paramétrique classique

- Le test paramétrique classique repose souvent sur l'hypothèse que le couple $(X, Y)^\top$ est de loi normale à deux dimensions:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \tau^2 & \rho\tau^2 \\ \rho\tau^2 & \tau^2 \end{pmatrix}\right)$$

- Ainsi, $E(\Delta) = \theta = \mu_Y - \mu_X$, et de plus:

$$\begin{aligned} \mathrm{var}(\Delta) &= \mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X, Y), \\ &= 2\tau^2 - 2\rho\tau^2, \\ &= 2\tau^2(1 - \rho). \end{aligned}$$

- Aussi, $\Delta \sim \mathcal{N}(\theta, 2\tau^2(1 - \rho))$.

Test d'hypothèse classique

- Dans le test classique, l'hypothèse nulle est $H_0 : \theta = 0$, versus l'une des trois hypothèses alternatives:

$$H_1 : \theta \neq 0; \; H_1 : \theta > 0; \; H_1 : \theta < 0;$$

- Le test-t usuel appliqué sur les différences est utilisé:

$$\bar{\Delta}_N = N^{-1} \sum_{i=1}^{N} \Delta_i, \; S_\Delta^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\Delta_i - \bar{\Delta}_N)^2;$$

- La statistique est:

$$t = \frac{\bar{\Delta}_N}{S_\Delta / \sqrt{N}} \sim t_{N-1},$$

sous H_0, et par exemple on rejette H_0 en faveur de $H_1 : \theta > 0$ lorsque $t \geq t_{N-1;1-\alpha}$.

Le test de permutation pour échantillons appariés

- On peut procéder comme dans le cas de deux échantillons.
- Les hypothèses sont cependant différentes.
- Dans le contexte d'échantillons appariés, il est présumé que sous H_0, la distribution de $\Delta = Y - X$, F_Δ, est symétrique:

$$\begin{aligned} F_\Delta(x) &= 1 - F_\Delta(-x), \\ P(\Delta \leq x) &= P(\Delta \geq -x), \ \forall x. \end{aligned}$$

- On remarque que l'hypothèse de distribution symétrique donne la relation $F_\Delta(x) + F_\Delta(-x) = 1$.

Statistique de test

- On considère la statistique $S = \sum_{i=1}^{N} \Delta_i$.
- On verra un résultat analogue à ce qui a été vu pour deux échantillons: un test basé sur S sera équivalent au test-t.
- On commence par écrire:

$$S = \sum_{i=1}^{N} |\Delta_i| \text{sign}(\Delta_i),$$

où la fonction $\text{sign}(\Delta_i)$ donne le signe de Δ_i:

$$\text{sign}(\Delta_i) = \begin{cases} 1, & \text{si } \Delta_i \geq 0, \\ -1, & \text{si } \Delta_i < 0. \end{cases}$$

Étude des signes

- On note que l'on peut reformuler $\text{sign}(\Delta_i)$ avec la fonction Heavyside:
$$\text{sign}(\Delta_i) = 2H(\Delta_i) - 1.$$
- Sous H_0, les variables aléatoires $H(\Delta_i)$ sont iid Bernoulli($\frac{1}{2}$).
- Ainsi, sous H_0,
$$\begin{aligned} P(\text{sign}(\Delta_i) = 1) &= P(H(\Delta_i) = 1) = \frac{1}{2}, \\ P(\text{sign}(\Delta_i) = -1) &= P(H(\Delta_i) = 0) = \frac{1}{2}, \end{aligned}$$
- Sous H_0, il y a autant de chances d'avoir un signe plus qu'un signe moins, et les $\text{sign}(\Delta_i)$, $i = 1, \ldots, N$ sont indépendantes.

Distribution de permutation

- Supposons que l'on a observé $\mathbf{\Delta} = (\Delta_1, \ldots, \Delta_N)^\top$.
- On pose $\mathbf{\Delta} = \boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)^\top$.
- Les signes associés sont:

$$\begin{aligned} \mathbf{s} &= (s_1, \ldots, s_N)^\top, \\ &= (\text{sign}(\Delta_1)), \ldots, \text{sign}(\Delta_N))^\top. \end{aligned}$$

- La distribution de permutation sera notée $\mathcal{P}$.
- La distribution sous $\mathcal{P}$ de la statistique S est alors:

$$\mathcal{P}\left(S = \sum_{i=1}^{N} |\delta_i| s_i\right) = \frac{1}{2^N},$$

où les $s_i \in \{-1, 1\}$.

Symétrie et distribution $\mathcal{P}$

- La distribution de S, sous la distribution de permutation $\mathcal{P}$, est symétrique par rapport à zéro.
- On va noter:
$$S \overset{\mathcal{L}_{\mathcal{P}}}{=} -S$$
- En effet, supposons que $x \in \text{support}(S)$. Ainsi, S est de la forme $s = \sum_{i=1}^{N} |\delta_i| s_i$, pour un certain vecteur $\mathbf{s} = (s_1, \ldots, s_N)^\top$.
- Cependant, le vecteur $(-s_1, \ldots, -s_N)^\top$ est également une possibilité.
- La valeur correspondante est $-x = \sum_{i=1}^{N} |\delta_i| s_i$.

Exemple d'un test unilatéral

- Supposons que l'on cherche à tester:

$$H_0 : \theta = 0,$$
$$H_1 : \theta > 0.$$

- Supposons que $q_\alpha^>$ est la plus petite valeur du support de S satisfaisant:

$$\mathcal{P}(S \geq q_\alpha^>) \leq \alpha.$$

- On rejettera H_0 lorsque S est trop grande, c'est-à-dire lorsque $S \geq q_\alpha^>$.

Le test de permutation vu comme un test conditionnel

- On note que la distribution de permutation peut s'écrire comme:
$$\mathcal{P}\left(S = \sum_{i=1}^{N} |\delta|_{(i)} s_i\right) = \frac{1}{2^N},$$
où le vecteur des signes $(s_1, \ldots, s_N)^\top \in \{-1, 1\}^N$.
- En effet, supposons que l'on a observé $(\delta_1, \ldots, \delta_N)^\top$
- On considère $\mathbf{R}^+ = (R_1^+, \ldots, R_N^+)^\top$ le vecteur constitué des rangs des $|\delta_i|$, $i = 1, \ldots, N$.
- On introduit également $\mathbf{D}^+ = (\mathbf{R}^+)^-$, le vecteur des positions des $|\delta|_{(i)}$ dans l'échantillon original des $|\delta_i|$.

Test de permutation (suite)

- On note que l'on peut écrire:

$$\begin{aligned}\sum_{i=1}^{N} |\delta_i| s_i &= \sum_{i=1}^{N} |\delta|_{(R_i^+)} s_{R_i^+} = \sum_{i=1}^{N} |\delta_{D_i^+}| s_{D_i^+}, \\ &= \sum_{i=1}^{N} |\delta|_{(i)} s_{D_i^+}.\end{aligned}$$

- Les deux ensembles suivants sont équivalents:

$$\left\{ \sum_{j=1}^{N} |\delta|_{(j)} s_j,\ (s_1, \ldots, s_N)^\top \in \{-1, 1\}^N \right\}$$

et

$$\left\{ \sum_{j=1}^{N} |\delta_j| s_j,\ (s_1, \ldots, s_N)^\top \in \{-1, 1\}^N \right\}$$

Test de permutation (suite)

- On considère la statistique

$$\begin{aligned} S &= \sum_{i=1}^{N} \Delta_i = \sum_{i=1}^{N} |\Delta_i| \mathrm{sign}(\Delta_i), \\ &= \sum_{i=1}^{N} |\Delta|_{R_i^+} \mathrm{sign}(\Delta_i), \end{aligned}$$

- On va conditionner sur $|\boldsymbol{\delta}|_{(\cdot)}$, c'est-à-dire que l'on va considérer $|\boldsymbol{\Delta}|_{(\cdot)} = (|\delta|_{(1)}, \ldots, |\delta|_{(N)})^\top$ comme étant fixé.
- On pose $H_i = H(\Delta_{D_i^+})$, $i = 1, \ldots, N$.
- On montre l'indépendance entre $(H_1, \ldots, H_N)^\top$ et $|\boldsymbol{\Delta}|_{(\cdot)}$.

Propositions

- Le résultat précédent repose sur une série de propositions intermédiaires.
- Résultat: Supposons que $\Delta_1, \ldots, \Delta_N$ sont iid d'une loi symétrique par rapport à zéro, de fonction de distribution F_Δ présumée continue. Alors $(H(\Delta_1), \ldots, H(\Delta_N))^\top$ et $|\mathbf{\Delta}|_{(\cdot)}$ sont indépendants.
- On rappelle que si la loi de Δ_i est symétrique ceci implique $P(\Delta_i \leq d) = P(\Delta_i \geq -d)$.
- Autrement dit:

$$F_\Delta(d) + F_\Delta(-d) = 1.$$

Corollaire

- On se rappelle que les rangs sont définis comme:

$$R_i^+ = \sum_{j=1}^{N} I\left(|\Delta_j| \leq |\Delta_i|\right)$$

- Le rang R_i^+ est une fonction des $|\Delta_i|$, $i = 1, \ldots, N$ ou de $\mathbf{\Delta}_{(\cdot)}$.
- On trouve donc que $\mathbf{R}^+ = (R_1^+, \ldots, R_N^+)^\top$ et $(H_1, \ldots, H_N)^\top$ sont indépendants.

Propositions (suite)

- Considérons à nouveau $\Delta_1, \ldots, \Delta_N$ iid avec F_Δ symétrique par rapport à zéro.
- Soient $|\Delta|_{(1)} < \ldots < |\Delta|_{(N)}$, les statistiques d'ordre associées aux $|\Delta_i|$.
- Posons $H_i = H(\Delta_{D_i^+})$, $i = 1, \ldots, N$.
- Résultat: Supposons que $\Delta_1, \ldots, \Delta_N$ sont iid d'une loi symétrique par rapport à zéro, de fonction de distribution F_Δ présumée continue. Alors les H_i sont indépendantes, chacune de loi de Bernoulli, c'est-à-dire que les H_i sont iid, $H_i \sim \text{Bernoulli}(1/2)$.

Résultat principal

- On rappelle que l'on cherche à montrer l'indépendance entre le vecteur aléatoire $(H_1, \ldots, H_N)^\top$ et $|\mathbf{\Delta}|_{(\cdot)}$.
- On montre directement que:

$$P(H_1 = h_1, \ldots, H_N = h_N \mid |\mathbf{\Delta}|_{(\cdot)} = |\boldsymbol{\delta}|_{(\cdot)}) = \frac{1}{2^N}$$

- Or les H_i sont iid Bernoulli, ce qui montre l'indépendance recherchée.

Ainsi, le test de permutation peut être vu comme un test conditionnel

- Revenons à $S = \sum_{i=1}^{N} |\Delta|_{R_i^+} \text{sign}(\Delta_i)$. Reprennant une argumentation comme dans le cas de deux échantillons:

$$
\begin{aligned}
\mathcal{L}(S \mid |\boldsymbol{\Delta}|_{(\cdot)} = |\boldsymbol{\delta}|_{(\cdot)}) &= \mathcal{L}(\sum_{i=1}^{N} |\Delta|_{(R_i^+)} \text{sign}(\Delta_i) \mid |\boldsymbol{\Delta}|_{(\cdot)} = |\boldsymbol{\delta}|_{(\cdot)}), \\
&= \mathcal{L}(\sum_{i=1}^{N} |\delta|_{(R_i^+)} \text{sign}(\Delta_i) \mid |\boldsymbol{\Delta}|_{(\cdot)} = |\boldsymbol{\delta}|_{(\cdot)}), \\
&= \mathcal{L}(\sum_{i=1}^{N} |\delta|_{(i)} \text{sign}(\Delta_{D_i^+}) \mid |\boldsymbol{\Delta}|_{(\cdot)} = |\boldsymbol{\delta}|_{(\cdot)}), \\
&= \mathcal{L}(\sum_{i=1}^{N} |\delta|_{(i)} (2H_i - 1) \mid |\boldsymbol{\Delta}|_{(\cdot)} = |\boldsymbol{\delta}|_{(\cdot)}), \\
&= \mathcal{L}(\sum_{i=1}^{N} |\delta|_{(i)} (2H_i - 1)),
\end{aligned}
$$

Loi conditionnelle de S

- La dernière ligne de l'argumentation précédente utilise l'indépendance entre $(H_1, \ldots, H_N)^\top$ et $|\mathbf{\Delta}|_{(\cdot)}$.
- Or le support de $\sum_{i=1}^N |\delta|_{(i)}(2H_i - 1)$ est:

$$\left\{ \sum_{i=1}^N |\delta|_{(i)} s_i, \quad (s_1, \ldots, s_N) \in \{-1, 1\}^N \right\}$$

- Ceci montre que:

$$P(S = \sum_{i=1}^N |\delta|_{(i)} s_i \mid |\mathbf{\Delta}|_{(\cdot)} = |\boldsymbol{\delta}|_{(\cdot)}) = \frac{1}{2^N}.$$

- Ceci est la même loi que:

$$\mathcal{P}(S = \sum_{i=1}^N |\delta|_{(i)} s_i) = \frac{1}{2^N}.$$

Liens avec le test-t pour échantillons appariés

- Considérons la statistique t que l'on réécrit comme suit:

$$t = \frac{\bar{\Delta}}{s_\Delta/\sqrt{N}} = \frac{\sqrt{\frac{N-1}{N}}\sum_{i=1}^{N}\Delta_i}{\left\{\sum_{i=1}^{N}\Delta_{(i)}^2 - N^{-1}(\sum_{i=1}^{N}\Delta_i)^2\right\}^{1/2}}$$

- Si on garde $|\mathbf{\Delta}|_{(\cdot)} = |\delta|_{(\cdot)}$ fixé, alors $\sum_{i=1}^{N}\Delta_{(i)}^2$ est aussi fixé. Or t est une fonction croissante de $\sum_{i=1}^{N}\Delta_i$.
- Ainsi, un test basé sur t sera équivalent à faire un test basé sur $S = \sum_{i=1}^{N}\Delta_i$.

Autre simplification

- Notons de plus que:

$$\begin{aligned} S &= \sum_{i=1}^{N} \Delta_i = \sum_{i=1}^{N} |\Delta_i| \text{sign}(\Delta_i), \\ &= \sum_{i=1}^{N} |\Delta_i|(2H(\Delta_i) - 1), \\ &= 2\sum_{i=1}^{N} |\Delta_i| H(\Delta_i) - \sum_{i=1}^{N} |\Delta_i|. \end{aligned}$$

- Posons $\tilde{S} = \sum_{i=1}^{N} |\Delta_i| H(\Delta_i)$.
- On peut conclure qu'effectuer un test de permutation basé sur S est alors équivalent à faire un test basé sur $\tilde{S}$.

Considérations numériques

- On note que

$$\begin{aligned}\tilde{S} &= \sum_{i=1}^{N} |\Delta_i| H(\Delta_i), \\ &= \sum_{i=1}^{N} |\Delta|_{(R_i^+)} H(\Delta_{R_i^+}), \\ &= \sum_{i=1}^{N} |\Delta|_{(i)} H(\Delta_{D_i^+}),\end{aligned}$$

- On aura donc que $\tilde{S} = \sum_{i=1}^{N} |\Delta|_{(i)} H_i$ avec H_i iid Bernoulli$(1/2)$ sous H_0.
- Le support de $\tilde{S}$ comprend 2^N valeurs.

Approximation de la distribution de permutation

- Supposons que l'on a observé les différences Δ_i, $i = 1, \ldots, N$.
- On dispose alors de $S = \sum_{i=1}^{N} \Delta_i$.
- On choisit au hasard B vecteurs de signes:

$$\mathbf{s}_b = (s_{1b}, \ldots, s_{Nb})^\top, \quad \text{avec } \mathbf{s}_b \in \{-1, 1\}^N.$$

- On peut alors calculer $S_b = \sum_{i=1}^{N} |\Delta|_{(i)} s_{ib}$, $b = 1, \ldots, B$.
- Puisque la distribution de S est symétrique, $-S_b$ est également possible.
- On obtient alors $2B$ valeurs possibles.
- Habituellement $2B = 1000$ devrait bien approximer la loi de S pour un niveau de 5%, et $2B = 5000$ devrait être suffisant pour un niveau de 1%.