On a vu les ANOVA sur régression linéaire et son fucken.

Hypothèse très forte sur le modèle : normalité des vecteurs d'erreur.

⇒ On connaît la loi des observation.

Problème général : X explique Y → indépendance !
liés ! → Y a les mêmes valeurs sur les modalités de X
→ Y évolue linéairement par rapport à X - .

Tests non paramétriques

Un test est dit non-paramétrique s'il peut être appliqué quelle que soit la distribution des observations. Puisque la distribution (normale ? exponentielle ? uniforme ? autre ?) n'est pas spécifiée, on n'a aucun paramètre à estimer. Il existe une multitude de méthodes et de tests non paramétriques, certains plutôt grossiers, d'autres beaucoup plus raffinés, qui permettent de tester plusieurs types d'hypothèses (équidistribution de deux échantillons, symétrie d'une distribution, indépendance, etc.) En voici quelques illustrations.

Aucune hypothèse sur les distributions

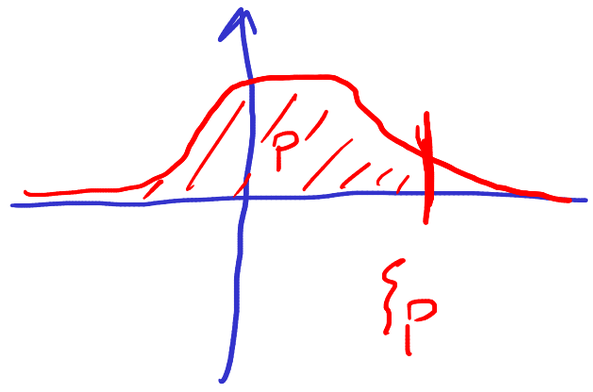
↳ on les considère continues

et c'est tout !

15 Intervalle pour les quantiles – Exemple introductif

Rappelons la notion de quantile. Soit X une variable alétaoire de type continu dont la densité est $f(x)$ et dont la fonction de répartition est $F(X)$. Soit p une proportion strictement entre 0 et 1. Supposons que $F(x) = p$ admette une unique solution en x. Cette solution est dite : "quantile d'ordre p". Cette racine (ou solution) est notée ξ_p. Ainsi,

$$P(X \leq \xi_p) = F(\xi_p) = p.$$



$X_1, \dots, X_n$ des observations.

$\downarrow$ on ordonne par ordre croissant

$Y_1 \leq Y_2 \leq \dots \leq Y_{n-1} \leq Y_n$.

On pose l'intervalle de confiance (dont les bornes sont des observations)

$$Y_i < \xi_p < Y_j.$$

Le niveau de confiance est $P(Y_i < \xi_p < Y_j)$

Événement $\{Y_i < \xi_p < Y_j\}$ Il y a au moins i observations inférieures à ξ_p ($Y_i < \xi_p$) et moins de j observations inférieures à ξ_p (sinon $Y_j \leq \xi_p$).

η compte le nombre de succès à l'expérience $X < \xi_p$, dont la proba de succès vaut $p(X<\xi_p) = p$. $\eta \sim \mathcal{B}(n,p)$

$$P(Y_i < \xi_p < Y_j) = P(i \leq \eta < j) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}.$$

Pour que la $i$ème statistique d'ordre soit inférieure à ξ_p il faut qu'il y ait au moins i valeurs observées inférieures à ξ_p. De plus, pour que la $j$ème statistique d'ordre soit supérieure à ξ_p, il faut qu'il y ait moins de j valeurs inférieures à ξ_p. Ainsi, si nous considérons qu'une valeur inférieure à ξ_p est un succès, alors, parmi les n essais indépendants, il doit y avoir entre i et $j-1$ succès pour que l'événement qui nous intéresse se réalise. Donc

$$P(Y_i < \xi_p < Y_j) = \sum_{k=i}^{j-1} \frac{n!}{k!(n-p)!} p^k (1-p)^{n-k}.$$

$Y_1 \leq \dots < Y_{10}$ $n = 10$ observations.

$\hookrightarrow$ $\xi_{0,75}$ $p = 0,75$.

Exemple.

On veut un intervalle de confiance pour le troisième quartile en observant un échantillon de taille 10. On a

$$P(Y_4 < \xi_{3/4} < Y_9) = \sum_{k=4}^{9-1} \frac{10!}{k!(10-k)!} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{n-k} \approx 0.9240.$$

$\hookrightarrow$ $P(4 \leq N_6 < 9)$

Donc $\xi_{0,75} \in]Y_4 ; Y_9[$ à 92,4% de confiance.

Ainsi, en utilisant Y_4 et Y_9, on obtient un intervalle de confiance à $92,40\%$. Les données sont dans le tableau suivant :

Un échantillon de taille 10 d'une population continue

i	1	2	3	4	5	6	7	8	9	10
X_i	10,48	14,43	0,82	4,03	4,74	4,70	4,21	0,53	14,73	0,91
Y_i	0,53	0,82	0,91	4,03	4,21	4,70	4,74	10,48	14,43	14,73

Un intervalle de confiance pour le 3e quartile est donc $]4,03; 14,43[$. Pour cet exemple, les auteurs ont échantillonné une population de loi $\mathcal{E}(5)$. On montre facilement que la valeur théorique de ce quartile est

$$Q3 = 10ln(2) = 6,93$$

16 Test d'équidistribution de deux échantillons

Comme souvent en statistiques, un problème ouvre une bibliothèque de tests. Nous proposons d'en détailler deux, le lecteur étant libre de se cultiver à volonté.

Nous avons beaucoup parlé du test du χ^2 en début de semestre ainsi qu'en début de formation tant pour tester l'adéquation d'une série observée à une série théorique que pour tester l'indépendance (qui revient à comparer une loi jointe observée à une loi jointe produit des marges). Puisque dans ce cas la distribution de la variable observée n'est pas supposée connue, nous pouvons parler de test non paramétrique.

H_0 = " X et Y sont équidistribuées ".

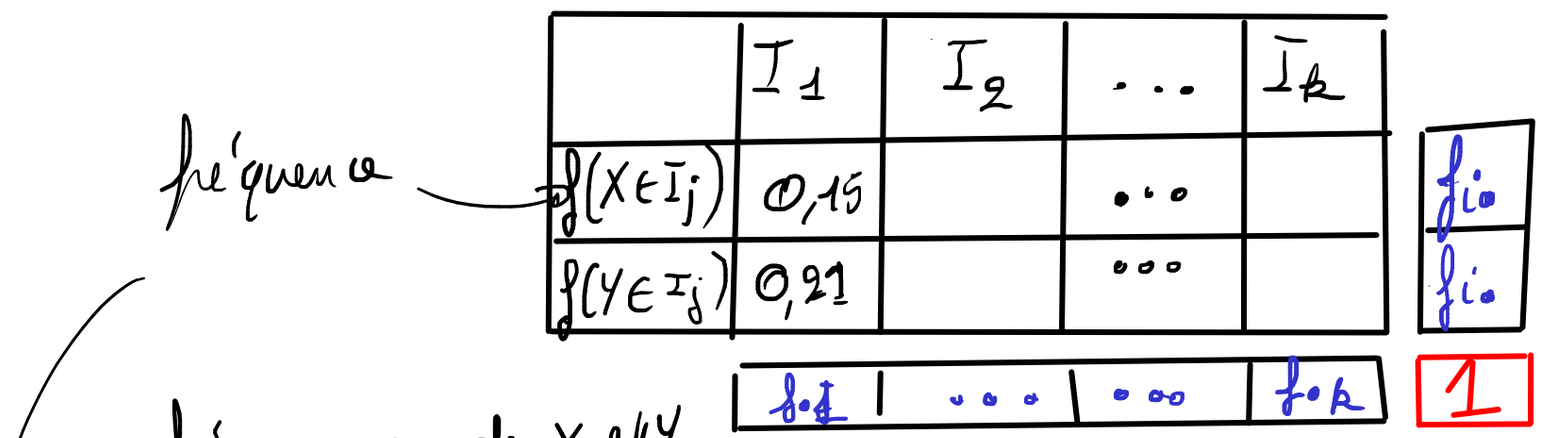
16.1 Test du χ^2... Encore !

Dire que deux variables X et Y sont équidistribuées est équivalent à dire que la distribution d'une variable est indépendante du fait qu'il s'agisse d'un X ou d'un Y.

En découpant l'intervalle $]-\infty, +\infty[$ en k tronçons $I_1, I_2, \cdots I_k$, on peut dénombrer

$$\begin{aligned} N_{1j} &= Card\{i : X_i \in I_j\} = \text{“nombre de } X \text{ dans } I_j \text{ ”}, \\ N_{2j} &= Card\{i : Y_i \in I_j\} = \text{“nombre de } Y \text{ dans } I_j \text{ ”}. \end{aligned}$$

On obtient donc un tableau $2 \times k$ et la suite se fait exactement comme s'il s'agissait d'un test d'indépendance.



$X_1, \dots, X_{n_1}$

$Y_1, \dots, Y_{n_2}$.

$n = n_1 + n_2$.

↳ fréquences de X et Y.

⇒ la somme des fréquences dans le tableau vaut 1 (et pas par ligne).

$\mathcal{H}_0$ = " Équi-répartition " = " le tableau dessus est une situation d'indépendance "

$$D\chi^2 = n \sum_{ij} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{(f_{i\bullet} f_{\bullet j})} \sim \mathcal{D}\chi^2(k-1).$$

→ équivalent de l'ANOVA non paramétrique pour facteur à 2 modalités.

16.2 Le test de Wilcoxon

Considérons deux échantillons $X_1, X_2, \cdots, X_{n_X}$ et $Y_1, Y_2, \cdots, Y_{n_Y}$. On veut tester l'hypothèse

$$\mathcal{H}0 : \text{“} F_X = F_Y''.$$

Pour plus de 2 modalité, le test s'appelle Kruskall-Wallis.

16.2.1 Le tri

Pour appliquer le test de Wilcoxon, on ordonne (disons, de la plus petite à la plus grande) l'ensemble des $n = n_X + n_Y$ observations. On obtient alors un mot formé de $n_X + n_Y$ lettres (n_X fois la lettre X et n_Y fois la lettre Y).

Exemple.

Si $X = (17.1, 14.5, 20.3, 8.3)$ et $Y = (5.2, 10.3, 12.4)$, on obtient le mot

$$YXYYYXXX$$

(la plus petite des 7 observations est un Y, la deuxième est un X, $\cdots$, les 3 plus grandes sont des X).

Puis on numérote : $\bar{Y}X\bar{Y}YYXXX$

Rangs 1 2 3 4 5 6 7

H_0 = "égalité de la somme des rangs de X et Y"

= "égalité des médianes" → d'où l'analogie avec l'ANOVA qui teste l'égalité des moyennes.

16.2.2 La variable

On choisit X ou Y de manière arbitraire.

La variable du test de Wilcoxon est

effectif de X

$$W = \sum_{i=1}^{n_X} R_i,$$

rangs de X.

où R_i désigne le rang de l'observation X_i parmi les $n = n_X + n_Y$ observations en ordre croissant.
$\mathcal{H}_0$ sera rejetée si W est significativement grand ou significativement petit.

Donnons la loi de W. Si $F_X = F_Y$ (i.e. $\mathcal{H}_0$) alors les rangs R_i suivent une loi $\mathcal{U}(\llbracket 1, n \rrbracket)$

$n = n_X + n_Y$.

$$E R_i = \frac{n+1}{2} \qquad \text{Var } R_i = \frac{n^2 - 1}{12}$$

$$E W = E \sum_{i=1}^{n_X} R_i = \sum_{i=1}^{n_X} E R_i = n_X \frac{(n+1)}{2}$$

$$\text{Var } W = \text{Var}\left(\sum_{i=1}^{n_X} R_i\right) = \sum_{i=1}^{n_X} \text{Var } R_i + \sum_{\substack{i,j=1 \\ i \neq j}}^{n_X} \text{Cov}(R_i, R_j)$$

⚠ pas indépendance

$\Big\}$ constante C

$$= n_X \frac{(n^2-1)}{12} + n_X(n_X-1)\, C$$

Supposons qu'il n'y ait pas de Y. $n_X = n$

donc tous les X sont au début et la variance est nulle.

$$\frac{n_X (n^2-1)}{12} + n_X(n_X-1)\, C = 0$$

$$\Leftrightarrow \frac{(n+1)(n-1)}{12} + (n-1)C = 0$$

$$\Leftrightarrow C = -\frac{n+1}{12}$$

On injecte dans la variance :

$$\text{Var } W = n_x \frac{(n^2-1)}{12} + n_x (n_x - 1)\left(\frac{-(n+1)}{12}\right)$$

$(n+1)(n-1)$

$$= \frac{n_x(n+1)}{12}\left[n - 1 - (n_x - 1)\right] \qquad n - n_x = n_y$$

$$= \frac{n_x n_y (n+1)}{12}$$

$W = \sum_i R_i$, $\quad \mathbb{E}W = n_x \frac{n+1}{2}$ $\quad \text{Var } W = \frac{n_x n_y (n+1)}{12}$

On suppose que n_x est relativement grand et on applique le théorème central limit (une version qui ne nécessite pas l'indépendance)

$$W \sim \mathcal{N}\left(n_x \frac{n+1}{2} ; \sqrt{\frac{n_y\, n_x(n+1)}{12}}\right)$$

En vertue du Théorème Central Limit, si n_X et n_Y sont tous deux grands, la statistique

$$W = \sum_{i=1}^{n_x} Ri$$

suivra asymptotiquement une loi normale

$$\mathcal{N}\left(\frac{n_X(n+1)}{2}; \sqrt{\frac{n_X n_Y(n+1)}{12}}\right).$$

16.2.3 Conclusion du test

1,96 pour $\alpha = 5\%$.

L'hypothèse $\mathcal{H}_0$ sera donc rejetée au seuil α si

$$\left| W - \frac{n_X(n+1)}{2} \right| > q_{\alpha/2} \sqrt{\frac{n_X n_Y (n+1)}{12}},$$

où $q_{\alpha/2}$ est le quantile théorique de la loi normale $\mathcal{N}(0;1)$:

$$P\big(\mathcal{N}(0;1) > q_{\alpha/2}\big) = \frac{\alpha}{2}.$$

Remarque :

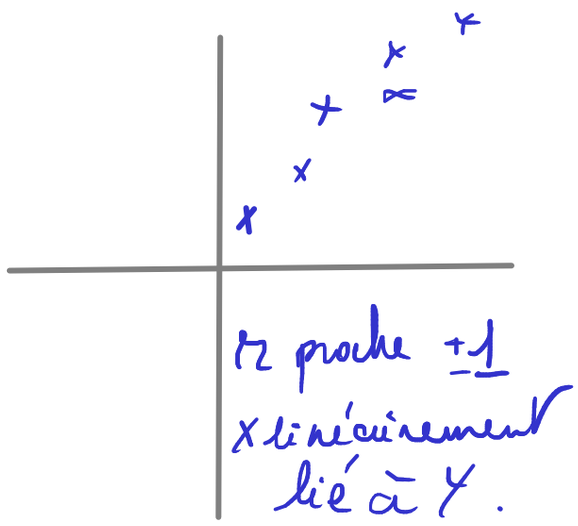
Le test de Wilcoxon est excellent pour détecter les différences de position (médiane, moyenne). Pour détecter les différences de dispersion, il ne vaut pas grand chose. Si, par exemple, on observe le mot

$$XXXXXXXXYYYYYYYYYYXXXXXXXX,$$

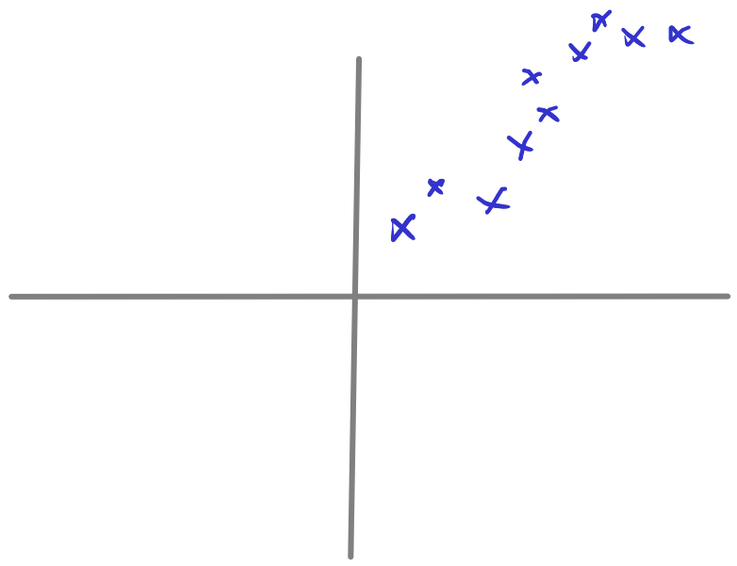
l'hypothèse d'équidistribution est évidemment fausse (tous les Y sont au centre et les X sont aux deux bouts) ; le test de Wilcoxon, pourtant, donnera une valeur de W tout-à-fait compatible avec l'hypothèse d'équidistribution.

17 Test d'indépendance–Test de Spearman

Nous avons déjà été amené à tester l'indépendance de deux variables avec le test du χ^2, en particulier pour les couples de variables qualitatives. Ce test reste le plus usité. On pourra également parler du test de Spearman basé sur la corrélation des rangs.



r proche ±1
X linéairement lié à Y.



r n'est plus proche de ±1
X et Y ne sont plus linéairement liées et pourtant, elles semblent liées.

On pourrait considérer un test basé sur le coefficient de corrélation échantillonnal r obtenu des n couples (X_i, Y_i) d'un couple de variables (X, Y). On a pu observer que la nullité de ce coefficient de corrélation n'entrainait pas à coup sûr l'indépendance des variables. En effet, le coefficient de corrélation r est pleinement aveugle face aux relations non linéaires du types $Y = X^2$.

Le coefficient de corrélation de rangs (noté R^*) est celui qu'on obtient en remplaçant simplement les observations X_i et Y_i par leurs rangs $R_{X(i)}$ et $R_{Y(i)}$. $R_{X(i)}$ est le rang obtenu par X_i dans l'échantillon X **ordonné** ; de même, $R_{Y(i)}$ est le rang de Y_i parmi les n valeurs de Y observées dans l'ordre. Le coefficient de corrélation de rangs est

$$R^* = corr(R_X, R_Y),$$

Idée, on ne fait plus la corélation sur les observations de X et Y mais sur leurs rangs :

$$R^* = \rho_{R_X R_Y}$$

R_X	1	4	2	5	3
X	1	2,5	1,2	2,8	2
Y	1,5	3,5	5	5,2	3
R_Y	1	3	4	5	2

et se calcule par la formule

$$R^* = \frac{\frac{1}{n}\sum_{i=1}^{n} R_X(i)R_Y(i) - \left(\frac{1}{n}\sum_{i=1}^{n} R_X(i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} R_Y(i)\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} R_X^2(i) - \left(\frac{1}{n}\sum_{i=1}^{n} R_X(i)\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} R_Y^2(i) - \left(\frac{1}{n}\sum_{i=1}^{n} R_Y(i)\right)^2\right)}}$$

$$= \frac{n\sum_{i=1}^{n} R_X(i)R_Y(i) - \left(\sum_{i=1}^{n} R_X(i)\right)\left(\sum_{i=1}^{n} R_Y(i)\right)}{\sqrt{\left(n\sum_{i=1}^{n} R_X^2(i) - \left(\sum_{i=1}^{n} R_X(i)\right)^2\right)\left(n\sum_{i=1}^{n} R_Y^2(i) - \left(\sum_{i=1}^{n} R_Y(i)\right)^2\right)}}.$$

On remarque toutefois que puisque les $R_X(i)$ (et les $R_Y(i)$) ne sont qu'une permutation des entiers de 1 à n, on a

$$\sum_{i=1}^{n} R_X(i) = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$
$$\sum_{i=1}^{n} R_X^2(i) = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\begin{aligned} n\sum_{i=1}^{n} R_X^2(i) - \left(\sum_{i=1}^{n} R_X(i)\right)^2 &= \frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4} \\ &= n^2(n+1)\left(\frac{2n+1}{6} - \frac{n+1}{4}\right) \\ &= \frac{n^2(n^2-1)}{12}. \end{aligned}$$

Le coefficient de corrélation de rangs devient donc

$$R^* = \frac{n\sum_{i=1}^{n} R_X(i)R_Y(i) - \frac{n^2(n+1)^2}{4}}{\frac{n^2(n^2-1)}{12}}$$
$$= \frac{12\sum_{i=1}^{n} R_X(i)R_Y(i)}{n(n^2-1)} - \frac{3(n+1)}{n-1}.$$

Posant $S^* = \sum_{i=1}^{n} R_X(i)R_Y(i)$, on observe que R^* est fonction linéaire de S^*. Noter que $R^* = AS^* + b$ où

$$a = \frac{12}{n(n^2-1)} \text{ et } b = -\frac{3(n+1)}{n-1}.$$

Il suffit donc de connaître la distribution de S^* pour connaître celle de R^*. En fait, on a

$$E[R^*|\mathcal{H}_0] = aE[S^*|\mathcal{H}_0] + b$$

et

$$Var[R^*|\mathcal{H}_0] = a^2 Var[S^*|\mathcal{H}_0].$$

En cette fin de chapitre, on se passera du détail des calculs. Si $\mathcal{H}_0$ est vraie, et si n est grand, La statistique S^* suivra approximativement une loi normale

TCL

$$\mathcal{N}\left(\frac{n^2(n+1)^2}{4}; \sqrt{\frac{n^2(n+1)(n^2-1)}{144}}\right).$$

On en déduit alors que R^* suit une loi normale (bien plus jolie)

$$\mathcal{N}\left(0, \sqrt{\frac{1}{n-1}}\right)$$

test asymptotique.

L'hypothèse $\mathcal{H}_0$ sera donc rejetée au seuil α si

$$|R^*| > q_{\alpha/2}\sqrt{\frac{1}{n-1}},$$

où $q_{\alpha/2}$ est le quantile théorique de la loi normale $\mathcal{N}(0;1)$:

$$P\big(\mathcal{N}(0;1) > q_{\alpha/2}\big) = \frac{\alpha}{2}.$$