Année Académique : 2022-2023

MASTER I

MATHÉMATIQUES ET APPLICATIONS :

INGÉNIERIE MATHEMATIQUES POUR LA SCIENCE DES DONNES

THÈME :

MESURER LE LIEN / L'INDEPENDANCE ENTRE DEUX VARIABLES, QUELS PARAMETRES ? QUELS TESTS ? POUR QUELLES CONCLUSIONS ?

Travail encadré de recherche

22/05/2023

Encadrée et dirigée par :	*Etudiants :*
Professeure Anne Gegout-Petit	Vanga Gustave Hermann Moulo Pascaline Kouda

Mesurer le lien / l'indépendance entre deux variables, quels paramètres ? quels tests ? pour quelles conclusions ?

PASCALINE KOUDA HERMANN GUSTAVE MOULO

Table des matières

Chapitre 1

Etude de liaisons entre deux variables quantitatives

L'étude de la liaison entre deux variables quantitatives se fait à partir de plusieurs indicateurs. Nous allons dans un présent abordé le premier indicateur qui est le coefficient de Pearson.

1.1 Corrélation de Pearson

1.1.1 Coéfficient de corrélation de pearson ou de corrélation linéaire

Le coéficient de corrélation linéaire ou encore coefficient de corrélation de pearson mesure l'intensité de liaison linéaire entre deux variables aleatoires. On la note ρ. Soit X et Y deux variables aléatoires, la formule de ρ est donnée par :

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Ce coéfficient est compris entre -1 et 1.

Ce coefficient depend étroitement de la liaison linéaire entre X et Y
Si nous disposons d'un échantillon de couple $(X_1, Y_1), \ldots, (X_n, Y_n)$ de taille n.
On definit :

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X_n})(Y_i - \bar{Y_n})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X_n})^2 \sum_{i=1}^{n}(Y_i - \bar{Y_n})^2}}$$

La valeur de r obtenue est une estimation de la corrélation entre deux variables continues dans la population. C'est la corrélation de l'échantillon

- $r \simeq -1$: forte corrélation linéaire négative entre X et Y ; il existe une liaison négative entre les deux variables(quand X augmente ,Y diminue)
- $r \simeq 0$: faible corrélation linéaire entre X et Y, il existe peut-être une liaison non linéaire entre X et Y
- $r \simeq 1$: forte corrélation linéaire positive entre X et Y ; il existe une liaison directe entre X et Y(quand X augmente Y augmente)

1.1.2 Test paramétrique sur le coefficient de corrélation

Conditions d'applications :
Le couple de variables (X, Y) doit se distribuer suivant une « *loi binormale* » qui est la généralisation de la loi normale à un couple de variables. Toute combinaison linéaire des deux variables est normale. En particulier, X et Y sont des variables normales. On doit vérifier ce postulat avant d'appliquer le test.

Hypothèse du test :
Sous le postulat de binormalité ,l'indépendance équivaut à $\rho = 0$. On teste alors
$H_0 : \rho = 0$ (indépendance)
$H_1 : \rho \neq 0$ ou $\rho > 0$ ou $\rho < 0$ (liaison,liaison positive,liaison négative)

Statistique du test :
La statistique est une variable aléatoire calculée sur les données de l'échantillon tiré au sort .sa valeur observée sur l'échantillon est un résumé des données permettant de choisir entre H_0 et H_1.
la statistique de test est :

$$t = \frac{r-\rho}{S_r}$$

suit la loi de student à $n-2$ dégré de liberté où

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$

sous $H_0, t = \frac{r}{S_r} \simeq Student(n-2)$,
Lorsque n est assez grand $t \simeq \mathcal{N}(0,1)$
Pour un seuil de signification α donné ,on rejette si $|t| \geq t_{\frac{\alpha}{2},n-2}$, où $t_{\frac{\alpha}{2},n-2}$ est le quantile de la loi de student à $n-2$

1.1.3 Lien entre la corrélation linéaire et la regression linéaire

La regression linéaire consiste à exprimer une variable Y en fonction d'une variable X avec une fonction affine. Ce qui donne la modélisation suivante :

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

- Y est la variable expliquée
- X est la variable explicative
- ϵ est le terme d'erreur aléatoire non observable
- β_0 et β_1 sont des paramètres à estimer

Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ n observations de X et Y nous pouvons donc écrire :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ pour } i = 1, \ldots, n$$

Comme nous l'avons vu en cours ,la problématique consiste à estimer les paramètres β_0 et β_1 ,en choisissant $\hat{\beta_0}$ et $\hat{\beta_1}$ de telle sorte que la distance Y_i et $\hat{\beta_0}+\hat{\beta_1}X_i$ soit minimale. Ainsi en utilisant la méthode des Moindres carres on trouve :

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X_n})(Y_i - \bar{Y_n})}{\sum_{i=1}^{n}(Y_i - \bar{Y_n})}$$

$$\hat{\beta}_0 = \bar{Y_n} - \hat{\beta}_1 \bar{X_n}$$

on definit $\hat{Y}$ l'estimation de Y. Ainsi pour n observations on a :

$$\hat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

On définit maintenant R^2 qui est égale :

$$R^2 = \frac{\sum_i^n (\hat{Y_i} - \bar{Y_n})^2}{\sum_i^n (Y_i - \bar{Y_n})^2}$$

Le R^2 s'appelle coefficient de determination qui sert à quantifier la part de variance de Y expliquée par la variable X. Ce coéfficient est le carré coefficient de corrélation de pearson . En effet

$$\mathrm{R}^2 = \frac{\sum_i^n (\hat{Y_i} - \bar{Y_n})^2}{\sum_i^n (Y_i - \bar{Y_n})^2}$$

$$\mathrm{R}^2 = \frac{\sum_i^n (\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y_n})^2}{\sum_i^n (Y_i - \bar{Y_n})^2}$$

$$\mathrm{R}^2 = \frac{\sum_i^n (\bar{Y_n} - \hat{\beta}_1 \bar{X_n} + \hat{\beta}_1 X_i - \bar{Y_n})^2}{\sum_i^n (Y_i - \bar{Y_n})^2}$$

$$\mathrm{R}^2 = \frac{\sum_i^n (-\hat{\beta}_1 \bar{X_n} + \hat{\beta}_1 X_i)^2}{\sum_i^n (Y_i - \bar{Y_n})^2}$$

$$\mathrm{R}^2 = \frac{\sum_i^n \hat{\beta_1}^2 (X_i - \bar{X_n})^2}{\sum_i^n (Y_i - \bar{Y_n})^2}$$

$$\mathrm{R}^2 = \frac{(\sum_{i=1}^n (X_i - \bar{X_n})(Y_i - \bar{Y_n}))^2 \sum_i^n (X_i - \bar{X_n})^2}{(\sum_{i=1}^n (X_i - \bar{X_n}))^2 \sum_i^n (Y_i - \bar{Y_n})^2}$$

$$\mathrm{R}^2 = \frac{(\sum_{i=1}^n (X_i - \bar{X_n})(Y_i - \bar{Y_n}))^2}{\sum_{i=1}^n (X_i - \bar{X_n})^2 \sum_{i=1}^n (Y_i - \bar{Y_n})^2}$$

On remarque le coefficient de determination R^2 est le carré de la corrélation r de l'échantillon.

1.2 Corrélation de spearman

1.2.1 Coefficient de corrélation de spearman

Le coéfficient de corrélation des rangs de spearman est une mesure de dépendance statistique non paramétrique entre deux variables. Elle consiste à trouver un coefficient de corrélation, non pas entre les valeurs prises par les deux variables mais entre les rangs de ces valeurs. On le nomme le plus souvent le ρ de spearman.
Ce coéfficient est aussi compris entre -1 et 1 et s'interprète de la manière que le coéfficient de corrélation de Pearson
Soient X, Y deux variables aléatoires et R_X et R_Y les variables de rangs de X et Y.
On a :

$$\rho(R_X, S_Y) = \frac{cov(R_X, S_Y)}{\sigma(R_X)\sigma(S_Y)}$$

Soient $(X_1, X_2, \ldots, X_n)$ et $(Y_1, Y_2, \ldots, Y_n)$ deux échantillons de taille n. On note R_{X_i} le rang que prend X_i par rapport aux autres valeurs de l'échantillon des X pour $i = 1, 2, \ldots, n$. Ainsi $R_{X_i} = 1$ si X_i est la plus petite valeur des X ,$R_{X_i} = 2$ si X_i est la deuxième petite valeur etc...jusqu'à $R_{X_i} = n$ si X_i est la plus grande valeur des X. De même , on note S_{Y_i} le rang de Y_i ,pour $i = 1, 2, \ldots, n$. On a :

$$r(R_X, S_Y) = \frac{\sum_{i=1}^{n}(R_{X_i} - \bar{R})(S_{Y_i} - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_{X_i} - \bar{R})^2 \sum_{i=1}^{n}(S_{Y_i} - \bar{S})^2}}$$

Où

$$\bar{R} = \sum_{i=1}^{n} R_{X_i} \quad \bar{S} = \sum_{i=1}^{n} S_{Y_i}$$

r peut s'écrire de la manière suivante :

$$r(R_X, S_Y) = 1 - \frac{\sum_i^n d_i^2}{n(n^2-1)}$$

Où $d_i = (R_{X_i}, S_{Y_i})$
Si plusieurs observations ont exacement la même valeur ,on attribue un rang moyen à ces observations . s'il y a beaucoup de rangs moyens ,il convient de faire une correction et de calculer :

$$r(R_X, S_Y) = \frac{S_x + S_y - 6\sum_i^n d_i^2}{2\sqrt{S_x . S_y}}$$

Où

$$S_x = \frac{n(n^2-1) - \sum_i^g (t_i^3 - t_i)}{12}$$

avec g le nombre de groupes de rangs moyens et t_i la taille du groupe de rang i pour l'échantillon des X et

$$S_y = \frac{n(n^2-1) - \sum_j^h (t_j^3 - t_j)}{12}$$

avec h le nombre de groupes de rangs moyens et t_j la taille du groupe de rang j pour l'échantillon des Y
S'il n'y a pas de rangs moyens ,les observations sont vues comme autant de groupes de taille 1. Par conséquent ,$g = h = n$ et $t_i = t_j = 1 \quad$ pour $\quad i, j = 1, \ldots, n$ et

$$S_x = S_y = \frac{n(n^2-1)}{12}$$

1.2.2 Test sur la significativité du coefficient de spearman

Hypothèse du test :
Sous le postulat de binormalité ,l'indépendance équivaut à $\rho = 0$. On teste alors
$H_0 : \rho = 0$ (indépendance)
$H_1 : \rho \neq 0$ ou $\rho > 0$ ou $\rho < 0$ (liaison,liaison positive,liaison négative)

Statistique du test :
La statistique est une variable aléatoire calculée sur les données de l'échantillon tiré au sort .sa valeur observée sur l'échantillon est un résumé des données permettant de choisir entre H_0 et H_1.
la statistique de test est :

$$T = \frac{r(R_X, S_Y) - \rho}{S_r(R_X, S_Y)}$$

suit la loi de student à $n-2$ dégré de liberté où

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$

sous $H_0, T = \frac{r}{S_r} \simeq Student(n-2)$,
Pour un seuil de signification α donné ,on rejette si $|t| \geq t_{\frac{\alpha}{2}, n-2}$, où $t_{\frac{\alpha}{2}, n-2}$ est le quantile de la loi de student à $n-2$

Remarque : Lorsque n est assez grand ,l'hypothèse de la normalité de deux variables aléatoires n'est

1.3 Coéfficient de corrélation de Kendall ou taux de Kendall et test associé

1.3.1 Coéfficient de corrélation de rangs de Kendall

Soient X et Y deux variables quantitatives continues observées sur un échantillon de taille n fournissant n paires d'observations $(X_1, X_2, \ldots, X_n)$ et $(Y_1, Y_2, \ldots, Y_n)$ On obtient une indication de la corrélation entre X et Y en ordonnant les valeurs X_i en ordre croissant et en comptant le nombre de valeurs Y_i correspondant ne satisfaisant pas cet ordre.
On note par$Q$le nombres d'inversions nécéssaires parmi les valeurs de Y pour obtenir le même ordre croissant que celui de X.
Comme il y a $\frac{n(n-1)}{2}$ paires distincts pouvant être formées, on a : $0 \leq Q \leq \frac{n(n-1)}{2}$; la valeur 0 est atteinte lorsque toutes les valeurs Y_i sont déja en croissant et la valeur $\frac{n(n-1)}{2}$ est atteinte lorsque les valeurs Y_i sont totalement dans l'ordre inverse des X_i chaque paire devant être échangée pour obtenir l'ordre recherché.
Le coéfficient de corrélation de rangs de Kendall, désigné par τ est défini par :

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

si toutes les paires sont dans l'ordre croissant ,

$$\tau = 1 - \frac{4.0}{n(n-1)} = 1$$

par contre ,si toutes les paires sont dans l'ordre inverse

$$\tau = 1 - \frac{4.\frac{n(n-1)}{2}}{n(n-1)} = -1$$

Il est possible de donner une définition équivalente du coéfficient des rangs de Kendall. Pour cela nous allons introduire les concepts suivants :

- Concordance
  Deux couples d'observations sont dites concordants si les deux membres de l'une des observations sont plus grands que le membre respectif de l'autre observation.

Par exemple $(0,9;1,3)$ et $(1,5;2,5)$ sont concordants car $0,9 < 1,5$ et $1,3 < 2,5$.

- discordance
  Deux couples d'observations sont dites discordants si les deux membres de l'une des observations sont d'ordre opposé par rapport au membre respectif de l'autre observation.

Par exemple $(0,9;2,5)$ et $(1,5;1,3)$ sont concordants car $0,9 < 1,5$ et $1,3 > 2,5$.

Soit N_c et N_d le nombre de paires d'observations respectivement concordantes et discordantes.
Deux paires pour lesquelles $X_i = X_j$ ou $Y_i = Y_j$ ne sont ni concordantes ni discordantes et ne sont donc comptabilisées ni dans N_c ni dans N_d.
Avec les concepts introduits, le coéfficient des rangs de kendall est défini par :

$$\tau = \frac{2(N_c - N_d)}{n(n-1)}$$

Lorsque qu'il n'y a pas de couples pour les lesquels $X_i = X_j$ ni $Y_i = Y_j$, les deux formulations de τ sont exactement les mêmes.

1.3.2 Test de la significativité

On se place le cas où il n'y a pas d'ex. On cherche à tester la significativité du τ .
S'il y a indépendance alors $\tau = 0 \rightarrow N_c = N_d$.
Hypothèses
H_0= les variables sont indépendantes
H_1= les variables sont dépendants

Statistiques de test
la statistique des test est : $z = \frac{\tau}{S_\tau}$ qui suit la loi normale centrée réduite
Où $S_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$

Règle de décision

Chapitre 2

Etude de liaisons entre deux variables qualitatives

2.1 Test de chi-deux d'indépendance

Tableau de contigence soient X et Y deux variables aléatoires qualitatives à k et p modalités respectivement.
on construit un tableau de contingence, c'est-à-dire un tableau dénombrant les modalités croisées des deux caractères X et Y.
Ce tableau aura donc k lignes (nombre de modalités de X) et p colonnes (nombres de modalités de Y). On lui adjoindra des marges où seront effectués les totaux en lignes (effectif de chaque modalité de X), les totaux en colonnes (effectif de chaque modalité de Y) et enfin le total général (nombre n d'individus étudiés).
Les différentes cases sont notées de façon abrégée à l'aide d'une variable N munie d'indices appropriés :

- N_{ij} : effectif de la case correspondant à la i^{me} ligne et la j^{ime} colonne du tableau, c'est-à-dire nombre d'individus ayant comme attribut la i^{me} modalité de X et la j^{ime} modalité de Y
- $N_{i.}$: somme de la ième ligne, c'est-à-dire nombre d'individus ayant comme attribut la i^{me} modalité de X
- $N_{.j}$: somme de la jème colonne, c'est-à-dire nombre d'individus ayant comme attribut la j^{ime} modalité de Y
- $N.. = n$: somme générale du tableau, c'est-à-dire nombre total d'individus étudiés

Ainsi on obtient le tableau suivant :

Tableau de contigence

	Y_1	...	Y_p	Total
X_1	n_{11}	...	n_{1p}	$n_{1.}$
...	...	...	...	...
X_k	n_{k1}	...	n_{kp}	$n_{k.}$
Total	$n_{.1}$	...	$n_{.p}$	$n..$

Conditions d'applications

- n doit etre assez grand c'est à dire $n \geq 30$
- $\frac{N_{i.}N_{.j}}{n} \geq 5$

Soit $p_{ij} = \mathbb{P}(X_i \cap Y_j)$ pour $i = 1, \ldots, k\,; \quad j = 1, \ldots, p$

Hypothèses :
$H_0 : p_{ij} = p_{i.}p_{.j}$
$H_1 : \exists (i,j) p_{ij} \neq p_{i.}p_{.j}$

Sous H_0,la statistique de test noté χ^2 suit la loi de chi-deux à $(k-1)(p-1)$ dégré de liberté.

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{(N_{ij} - \frac{N_{i.}N_{.j}}{n})^2}{\frac{N_{i.}N_{.j}}{n}} \simeq \chi^2[(k-1)(p-1)]$$

on rejette l'hypothèse nulle lorsque χ^2 est supérieur au quantile de la loi de chi-deux à $((k-1)(p-1))$ au seuil α

2.2 Test d'exact de Fisher

Conditions d'applications : Ce test est à utiliser pour tester l'indépendance de deux variables qualitatives pour des échantillons petits lorsque après regroupements ultimes il reste un tableau 22 i.e $\frac{N_{i.}N_{.j}}{n} \geq 5$ pour tout (i,j).
On dispose donc du tableau suivant

	Modalité Y_1 de Y	Modalité Y_2 de Y	
Modalité X_1 de X	A		$a+b$
Modalité X_2 de X			$c+d$
	$a+c$	$b+d$	N

Avec $N = a+c+b+d$
Sous H_0, $A \simeq Hypergometrique(N, n = a+b, p = \frac{a+c}{N})$ et on a :

$$\mathbf{P}(X = k) = \frac{\binom{a+c}{k}\binom{b+d}{n-k}}{\binom{N}{a+b}}$$

Dans la pratique, on ne va pas calculer la la zone de rejet mais directement calcluer la p-valeur. Celle-ci va etre égale à la probabilité d'avoir sous l'hypothese d'indépendance, le tableau observé ou quelque chose qui s'éloigne encore plus de l'hypothese d'indépendance. On aura soit $\mathbf{P}(X \geq a)$ ou $\mathbf{P}(X \leq a)$. On compare ensuite au risque α choisi.

2.3 Test de MacNemar

(pas encore rédiger).....

Chapitre 3

Etude de liaisons entre une variable qualitative et une variable quantitative

Nous désirons mesurer la liaison entre une variable qualitative et une variable quantitative à travers des tests statistiques.
Soient X la variable qualitative et Y la variable quantitative. Deux cas se présente à nous. Nous avons donc le cas où la variable X à deux modalités et le deuxième cas où X à $K > 2$ modalités.
On se place dans le cadre où la variable X à deux modalités. Si on extrait un échantillon $(Y_{1,1}, Y_{2,1}, ..., Y_{n_1,1})$ correspondant à la prémière modalité $X_i = 1$ et un autre échantillon $(Y_{1,2}, Y_{2,2}, ..., Y_{n_2,2})$ correspondant à la deuxième modalité $X_i = 2$ de la variable qualitative X, s'il n'y a pas de lien entre ces deux échantillons, ils doivent avoir la même distribution.
Et un moyen de vérifier s'ils ont la même distribution est de régarder s'ils ont la même moyenne.
Ce qui nous amène à introduit le test de Student qui est un test de comparaison de moyennes de deux échantillons indépendants.

3.1 Test de Student

Le test de Student est l'un des tests statistiques le plus utilisé pour comparer les moyennes de deux groupes indépendants.
Formulation : Soient X la variable qualitative à deux modalités et $(Y_{1,1}, Y_{2,1}, ..., Y_{n_1,1})$ et $(Y_{1,2}, Y_{2,2}, ..., Y_{n_2,2})$ les deux échantillons de la variable quantitative Y de tailles respectives n_1 et n_2.
Supposons que les deux échantillons sont indépendants de loi respectives $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$ avec $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Pour $j = 1, 2$, on note $\overline{Y_j}$ la moyenne empirique et S_j^2 la variance empirique corrigée de l'échantillon j. On a pour le premier échantillon :

$$\overline{Y_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1,i} \text{ et } S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1,i} - \overline{Y_1})^2$$

Hypothèses

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

La statistique de test T est donnée par la formule suivante :

$$T = \frac{(\overline{Y_1}-\overline{Y_2})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

où

$$S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

Sous H_0, la statistique de test T suit la loi de Student à n_1+n_2-2 degrés de liberté. Soit $t_{\frac{\alpha}{2}}$ le quantile $q_{1-\frac{\alpha}{2}}$ de la loi de Student à n_1+n_2-2 degrés de liberté au seuil α
Preuve Comme, les deux échantillons sont indépendants de loi respectives $\mathcal{N}(\mu_1,\sigma_1^2)$ et $\mathcal{N}(\mu_2,\sigma_2^2)$ avec $\sigma_1^2=\sigma_2^2=\sigma^2$ on a $\overline{Y_1}-\overline{Y_2} \simeq \mathcal{N}(\mu_1-\mu_2, \frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2})$ Il vient d'après le théoreme Central Limite,

$$\frac{(\overline{Y_1}-\overline{Y_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}} \simeq \mathcal{N}(0,1)$$

S_1^2 et S_2^2 sont des estimateurs de σ^2 donc S_p^2 est aussi un estimateur de σ^2. De plus d'après le Cochran, $\frac{(n_1-1)S_1^2}{\sigma^2} \simeq \chi^2_{n_1-1}$ et $\frac{(n_1-1)S_2^2}{\sigma^2} \simeq \chi^2_{n_2-1}$ et donc $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \simeq \chi^2_{n_1+n_2-2}$. D'où

$$T = \frac{(\overline{Y_1}-\overline{Y_2})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \simeq \chi^2_{n_1+n_2-2}$$

On rejette H_0 si $|T|$ est supérieure à $t_{\frac{\alpha}{2}}$
Lorsque les deux d'échantillons ont des variances inégales, le test de Welch peut être utilisé.

3.2 Test de Welch

Le test de Welch est une adaptation du test de Student. Il est utilisé pour comparer deux echantillons lorsque l'égalité des variances n'est pas assurée.
Formulation Soient X la variable qualitative à deux modalités et $(Y_{1,1}, Y_{2,1}, ..., Y_{n_1,1})$ et $(Y_{1,2}, Y_{2,2}, ..., Y_{n_2,2})$ les deux échantillons de la variable quantitative Y de tailles respectives n_1 et n_2.
Supposons que les deux échantillons sont indépendants de loi respectives $\mathcal{N}(\mu_1,\sigma_1^2)$ et $\mathcal{N}(\mu_2,\sigma_2^2)$ avec $\sigma_1^2 \neq \sigma_2^2$.
La statistique de test T_w donnée par :

$$T_w = \frac{\overline{Y_1}-\overline{Y_2}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

Hypothese

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

Si n_1, $n_2 \geq 30$, la statistique de test T_w suit approximativement la loi normale centrée réduite. Soit $z_{\frac{\alpha}{2}}$ le quantile $q_{1-\frac{\alpha}{2}}$ de la loi normale au seuil α. On rejette H_0 si $|T_w|$ est supérieure à $z_{\frac{\alpha}{2}}$
Sinon T_w suit la loi de Student sous H_0 mais le nombre de degrés de liberté est plus difficile à déterminer.

Comme dans ce qui précède, nous avons regarder la moyenne, à présent nous allons nous interesser à autre chose que la moyenne donc les distributions. Et le test de wilcoxon y répond.

3.3 Test de Wilcoxon

Le test de Wilcoxon (dont l'équivalent est le test de Mann-Whitney) ou test des rangs est un test non-paramétrique qui teste l'égalité des distributions des deux séries indépendantes de valeurs à comparer. Le principe du test est le suivant : on trie les valeurs des deux séries mises en commun dans le sens croissant, on attribue le rang 1 à la plus petite valeur, le rang 2 à la valeur suivante et ainsi de suite. On calcule ensuite le score de chacune des séries en sommant les rangs obtenus pour chacune d'elle. Sous l'hypothèse d'absolue continuité de la loi sous-jacente, il n'y a pas d'ex aequo.
A l'aide d'une table adéquate, on décide si ces scores sont compatibles avec l'hypothèse H_0 d'égalité des distributions. Le test de Mann-Whithney/Wilcoxon s'applique sans condition d'application.

Formulation On se place dans le cadre que précédemment
Soient X la variable qualitative à deux modalités et $(Y_{1,1}, Y_{2,1}, ..., Y_{n_1,1})$ et $(Y_{1,2}, Y_{2,2}, ..., Y_{n_2,2})$ les deux échantillons de la variable quantitative Y de tailles respectives n_1 et n_2 correspondants respectivement aux deux modalités de la variable qualitative X. Supposons que les deux échantillons sont indépendants.

Hypothèses :
- H_0 : les deux échantillons ont la même distibution
- H_1 : Cas bilatéral : les deux échantillons n'ont pas la même distribution

La statistique de test est le score de l'échantillon qui a le plus petit effectif et est donné par :

$$S = \sum_{i=1}^{n_1} R_i$$

où R_i le rang attribué à X_i et en supposant $n_1 \leq n_2$.
- (a) : Cas où $min(n_1, n_2) = n_1 \leq 10$, S ne suit pas une loi usuelle, mais on trouve les probabilités correspondantes dans la table de Mann-Whitney/Wilcoxon.
- (b) : Cas où $min(n_1, n_2) = n_1 \geq 10$, S suit approximativement sous H_0, une loi normale i.e :

$$S \simeq \mathcal{N}(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12})$$

En plus du test de Wilcoxon, on peut aussi utiliser le test de non paramétrique de Kolmogorov-Smirnov lorsque nous avons deux modalités et que l'hypothese sur la normalité n'est pas assurée.

3.4 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test d'adéquation non paramétrique qui vise à déterminer si les fonctions de répartition de deux populations sont identiques. Il est utilisé lorsqu'on est en présence de deux échantillons provenant de deux populations pouvant etre différentes.

Formulation Soient X la variable qualitative à deux modalités et $(Y_{1,1}, Y_{2,1}, ..., Y_{n_1,1})$ et $(Y_{1,2}, Y_{2,2}, ..., Y_{n_2,2})$ les deux échantillons de la variable quantitative Y de tailles respectives n_1 et n_2 correspondants respectivement aux deux modalités de la variable qualitative X. Supposons que les deux échantillons sont indépendants.

On note respectivement F_1 et F_2 leur fonction de répartiton.

Hypotheses $H_0 : F_1 = F_2$ pour tout x,
$H_1 : F_1 \neq F_2$ pour au moins une valeur de x
On note $\hat{F}^1_{n_1}(t)$ et $\hat{F}^2_{n_2}(t)$ les fonctions de répartitions empiriques des deux échantillons.
Le test statistique est défini par :

$$D_{n_1,n_2} = \sup_t |\hat{F}^1_{n_1}(t) - \hat{F}^2_{n_2}(t)|$$

avec

$$\hat{F}^1_{n_1}(t) = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{1}_{Y_{i,1} \leq t}$$

On rejette H_0 au seuil α si $D_{n_1,n_2} > t^{1-\alpha}_{n_1,n_2}$ ou $t^{1-\alpha}_{n_1,n_2}$ est la valeur de la table de Smirnov avec pour paramètre (n_1, n_2)

On se place à présent dans le cadre où la variable qualitative X à $K > 2$ modalités.
Si on extrait un échantillon $(Y_{1,1}, Y_{2,1}, ..., Y_{n_1,1})$ correspondant à la prémière modalité $X_i = 1$ et un autre échantillon $(Y_{1,2}, Y_{2,2}, ..., Y_{n_2,2})$ correspondant à la deuxième modalité $X_i = 2$ jusqu'à $(Y_{1,K}, Y_{2,K}, ..., Y_{n_K,K})$ correspondant à la modalité $X_i = K$de la variable qualitative X

Si l'hypothèse d'homocédasticité est vérifiée on fait une ANOVA dans le cas contraire, on utilise le test de Kruskal Wallis qui est une extension du test de Mann-Withney/Wilcoxon.

3.5 Analyse de la variance ANOVA

L'analyse de variance pour un facteur à K niveaux (ou K traitements) est une technique permettant de déterminer s'il existe une différence significative entre les K traitements (populations). Il est utilisé lorsque l'hypothese de normalité est vérifiée.

Le principe de l'analvse de variance à un facteur est, à partir d'échantillons tirés de ces populations, de comparer la variabilité à l'intérieur de chaque échantillon avec la variabilité entre les échantillons. La source de variation que l'on appelle erreur correspond donc à la variabilité à l'intérieur des échantillons, celle que l'on appelle effet correspond à la variabilité entre les échantillons.

Conditions d'application : Soient Y une variable quantitative d'intéret et une variable qualitative X avec K modalités.
Le modele linéaire est donné par : $Y_{i,k} = \mu_k + \epsilon_{i,k}$ μ_k est l'esperance de Y pour l'individu qui a la modalité K

- $\epsilon_{i,k}$ iid et suit la loi normale $\mathcal{N}(0, \sigma^2)$
- Nous avons une homoscédasticité i.e tous les bruits sont la meme variance.
- $Y_{i,k} \simeq \mathcal{N}(0, \sigma^2)$

Hypotheses

$$H_0 : \mu_1 = \mu_2 = ... = \mu_K = \mu \text{ vs } H_1 : \exists i, j \text{ tels que } \mu_i \neq \mu_j$$

L'objectif à présent est d'estimer μ_k et σ^2 et de trouver la statistique de test et sa distribution sur H_0.

On estime μ_k par $\hat{\mu}_k = \bar{Y}_{.k} = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{i,k}$

$$\epsilon_{i,k} = Y_{i,k} - \bar{Y}_{.k} \text{ le résidus}$$

σ^2 est estimer par $\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{i,k} - \bar{Y}_{.k})^2$

Décomposition de la variabilité : La variabilité des n observations (K échantillons) est mésurée par la somme totale des carrés des écarts qui est définie par :

$$SC_T = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{i,k} - \bar{Y}_{..})^2 \text{ avec } \bar{Y}_{..} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_i, k$$
$$SC_T = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{i,k} - \bar{Y}_{.k})^2 + \sum_{i=1}^{n_k} n_k (\bar{Y}_{.k} - \bar{Y}_{..})^2$$

$SC_T = SC_R + SC_B$ avec

$$SC_R = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{i,k} - \bar{Y}_{.k})^2$$

est la somme des carrés associée à l'interieur des groupes

$$SC_B = \sum_{i=1}^{n_k} n_k (\bar{Y}_{.k} - \bar{Y}_{..})^2$$

est la somme des carrés entre les groupes.
On a :

$$\frac{n-K}{\sigma^2} \hat{\sigma}^2 = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{i,k} - \bar{Y}_{.k})^2}{\sigma^2} \simeq \chi^2(n-K)$$

Sous H_0, SC_R et SC_B sont indépendant et par le théoreme de Cochran on a :

$$\frac{SC_R/(n-K)}{\sigma^2} \simeq \chi^2(n-K)$$
$$\frac{SC_B/(K-1)}{\sigma^2} \simeq \chi^2(K-1)$$

Soit F la statistique de test. On a :

$$F = \frac{SC_B/(K-1)}{SC_R/(n-K)}$$

Sous H_0 F suit la loi de Fisher Snedecor à (K-1,n-K) dégré de liberté.
On rejette H_0 si $F > f^{1-\alpha}_{K-1,n-K}$ le quantile de la loi de Fisher au seuil α. On conclut dans ce cas qu'il existe une différence significative entre les moyennes des échantillons.

Lorsque nous avons plus de deux de modalités et que l'hypothese de normalité n'est pas vérifiée, on utilise le test non paramétrique le Test de Kruskal Wallis.

3.6 Test de Kruskal Wallis

Le but est déterminé si toutes les populations sont identiques ou si au moins une des populations tend à fournir des observations différentes des autres.

Formulation Soit K échantillons avec $(Y_{1,j}, Y_{2,j}, ..., Y_{n_K,j})$ l'échantillon j, $j = 1, 2, ..., n_K$, on désigne donc par n_j la taille de l'echantillon j pour $j = 1, ..., K$. On a $N = \sum_{j=1}^{K} n_j$
On classe les N observations par ordre croissant sans tenir compte de l'appartenance aux échantillons.
Soit $Y_{i,j}$ la i-eme observation de l'échantillon j et $R_{i,j}$ le rang attribué à $Y_{i,j}$. Soit

$$R_j = \sum_{i=1}^{n_j} R_{i,j} \text{ avec } j = 1, ..., K$$

la somme des rangs attribués aux observations de l'échantillon j. Soit $\bar{R}$ la moyenne globale des rangs et $\bar{R_j}$ la moyenne des rangs pour les observations de l'échnatillon j, la statistique de Kruskal-Wallis est définie de la manière suivante :

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{K} n_k(\bar{R_k} - \bar{R})^2$$

C'est bien l'expression d'une variabilité inter-classes i.e la dispersion autour de la moyenne globale. Or nous savons que $\bar{R} = \frac{N+1}{2}$ Il est possible de simplifer l'expression ci-dessus par la formule suivante :

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{K} \frac{S_k^2}{n_k} - 3(N+1)$$

où S_k est la somme des rangs des individus appartenant à l'échantillon k.

Hypothèses :
- H_0 : les K échantillons ont la même distibution
- H_1 : Au moins un des échantillons n'a pas la même distribution que les autres.