Chapitre 11 – Régression linéaire simple

Analyse de corrélation linéaire (R)

- permet déterminer l'intensité de la liaison linéaire entre 2 VAR
- utilise R linéaire pour mesure l'intensité de la liaison,

$$R=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{n\sum X_iY_i-(\sum X_i)(\sum Y_i)}{\sqrt{n\sum {X_i}^2-(\sum X_i)^2}\sqrt{n\sum {Y_i}^2-(\sum Y_i)^2}}=$$

$$\frac{\sum_{i=1}^{n}X_iY_i-\frac{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n}}{\sqrt{\sum_{i=1}^{n}{X_i}^2-\frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n}{Y_i}^2-\frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n}}}$$

Coefficient de corrélation (LINÉAIRE)

Nombre compris entre $-1 \le R \le 1$

Analyse de régression

- On veut déterminer la relation statistique entre deux variables Y et X (c'est l'analyse de régression)
- En général, la variable X est contrôlable
- La 1e étape d'une analyse de régression consiste à faire un graphique de Y en fonction de X

Variables dépendante et indépendante

Variable dépendante Y: variable expliquée
Variable indépendante X: variable explicative

Régression linéaire simple

Permet de déterminer une équation reliant la variable dépendante Y à la variable indépendante X, de déterminer la validité du modèle obtenu, de faire des prévisions

La droite de régression

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, β_0 et β_1 sont les paramètres à estimer
ROUGE : Variable expliquée (variable aléatoire)
Terme constant pour une valeur donnée de X_i (variable explicative)
ORANGE : Facteur aléatoire

Composantes du modèle

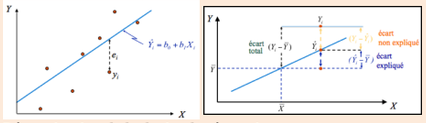
$Y_i = composante\ non\ aléatoire + composante\ aléatoire$
Composante non aléatoire : $E(Y_i/X_i) = \beta_0 + \beta_1 X_i$
Composante aléatoire : $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$

Hypothèses fondamentales du modèle

Les ε_i sont des VA ind. (normalement distribuées, d'esp math $E(\varepsilon_i) = 0$ et VAR constante $Var(\varepsilon_i) = {\sigma_\varepsilon}^2$)pour toutes valeurs de X_i

Estimation de la droite de régression

Plusieurs droites peuvent être ajustées aux données obtenues
Par la méthode des moindres carrés, on peut déterminer celle qui minimise la somme des carrés des erreurs --- $E(Y_i) = \beta_0 + \beta_1 X_i$ avec un estimateurs, $\hat{Y}_i = b_0 + b_1 X_i$



Détermination de la droite de régression

- Les résidus : $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 X_i)$
- La somme des carrés des erreurs : $SC_{rés} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- b_0 *et* b_1 sont les coefficients de régression (b_0 = l'ordonnée à l'origine & b_1 = la pente de la droite de régression empirique)

Méthode des moindres carrés

- On veut minimiser la somme des carrés des erreurs, $SC_{rés}$
- $Min\ SC_{rés} = \sum_{i=1}^{n} e^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$
- En solutionnant les équations précédentes, on obtient :
- $b_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{n\sum x_iy_i - \sum x_i \sum y_i}{n\sum {x_i}^2-(\sum x_i)^2}$ & $b_0 = \bar{y} - b_1\bar{x}$

Droite de régression empirique

$\hat{Y} = b_0 + b_1 X_i$, puisque $b_0 = \bar{Y} - b_1\bar{X}$ – alors en substituant b_0, on obtient $\hat{Y}_i = \bar{Y} + b_1(X_i - \bar{X})$

Analyse de la variance

Dans quelle mesure la droite de régression est-elle utile pour expliquer la variation des observations Y_i

- Variation expliquée par la régression
- Variation non expliquée ou résiduelle

Décomposition de la variation totale (voir graphique)

$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$
Écart total = écart expliqué par la droite + écart non expliqué

Somme des carrés et calcul de somme des carrés

Variation total = variation ex. par droite + variation non expliquée

$SCT = SCR + SC_{rés} = \sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$

- $SCT = \sum(Y_i - \bar{Y})^2 = \sum {Y_i}^2 - \frac{(\sum Y_i)^2}{n}$
- $SCR = \sum(\hat{Y}_i - \bar{Y})^2 = {b_1}^2\left[\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}\right]$
- $SC_{rés} = \sum(Y_i - \hat{Y}_i)^2 = \sum {Y_i}^2 - b_0\sum Y_i - b_1\sum X_iY_i$

Coefficient de détermination simple r^2

La qualité de l'ajustement du modèle aux données peut être mesurée grâce au coefficient de détermination simple r^2
Ce coefficient donne contribution de variable explicative ds l'équation de régression pour expliquer les variations de la variable dépendante

$$r^2 = \frac{variation\ expliquée}{variation\ totale} = \frac{SCR}{SCT} = \frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2},\ 0 \le r^2 \le 1$$

Coefficient de corrélation linéaire simple

Le signe du coefficient de corrélation est le même que celui du coefficient de régression b_1, $r = \pm\sqrt{r^2}, -1 \le r \le 1$

Variation totale inexpliquée

$$(1-r^2) = 1 - \frac{variation\ expliquée}{variation\ totale} = \frac{variation\ résiduelle}{variation\ totale} = 1 - \frac{SCR}{SCT} = \frac{SC_{rés}}{SCT}$$

Les sommes des carrés et leurs degrés de liberté

Variation	Somme des carrés	Degré de liberté
Régression	$SCR = \sum(\hat{Y}_i - \bar{Y})^2$	1
Résiduelle	$SC_{rés} = \sum(Y_i - \hat{Y}_i)^2$	(n-2)
Totale	$SCT = \sum(Y_i - \bar{Y})^2$	(n-1)

Carrés moyens ou variances

$CMR = \frac{SCR}{1} = \frac{\sum(\hat{Y}_i-\bar{Y})^2}{1}$ $\quad CM_{rés} = \frac{SC_{rés}}{n-2} = \frac{\sum(Y_i-\hat{Y}_i)^2}{n-2} = {S_e}^2$

Espérance des carrés moyens

$E(CM_{rés}) = {\sigma_E}^2$ $\quad E(CMR) = {\sigma_E}^2 + {\beta_1}^2\sum(X_i - \bar{X})^2$ si $\beta_1 = 0$, alors $E(CMR) = E(CM_{rés})$ et si $\beta_1 \neq 0$, alors $E(CMR) > E(CM_{rés})$

Tableau d'analyse de variance

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens	F
Due à la régression	SCR	1	CMR	$CMR/{S_e}^2$
Résiduelle	$SC_{rés}$	n-2	${S_e}^2$	
Total	SCT	n-1		

Test sur β_1 : test de Fisher **les valeurs critique sont tjrs ds tables

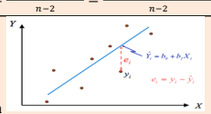
- Si $\beta_1 = 0$, alors $CMR/CM_{rés} = 1$
- $F = \frac{CMR}{CM_{rés}} = \frac{SCR/{\sigma_E}^2/1}{SC_{rés}/{\sigma_E}^2/n-2} = \frac{SCR/1}{SC_{rés}/n-2} = F_\alpha(1, n-2)$
- $H_0 : \beta_1 = 0,\ [E(CMR) = E(CM_{rés})]$
  $H_1 : \beta_1 \neq 0,\ [E(CMR) > E(CM_{rés})]$
- Seuil de signification α
- Si H_0 est vraie alors $F = \frac{CMR}{CM_{rés}} \in F(1, n-2)$
- Règle de décision : rejeter H_0 si $F > F_{\alpha;1,n-2}$

Résidus et mesure de distribution ${\sigma_e}^2$

- Veut obtenir estimation de dist. des Y_i autour de droite régression
- $Var(Y_i/X_i) = {\sigma_e}^2 = Var(\varepsilon_i)$
- Les résidus : $e_i = y_i - \hat{y}_i$ = permettent de vérifier l'hypothèse de normalité des erreurs et l'homogénéité de la variance des erreurs

Estimation de la variance σ^2

$${S_e}^2 = \frac{\sum(Y_i-\hat{Y}_i)^2}{n-2} = \frac{SC_{rés}}{n-2} = \frac{\sum {e_i}^2}{n-2} = \frac{\sum(Y_i-b_0-b_1X_i)^2}{n-2} = \frac{\sum {Y_i}^2 - b_0\sum Y_i - b_1\sum X_iY_i}{n-2}$$



Résidus et droite de régression

Analyse graphique des résidus

L'analyse des résidus permet de vérifier que : $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = {\sigma_\varepsilon}^2$
- les ε_i sont normalement distribués & ne sont pas autos corrélées
- L'analyse des résidus peut se faire graphiquement
- Les résidus devraient se distribuer aléatoirement autour de zéro(0)
- Si l'analyse du graphique des résidus en fonction de la variable dépendante révèle une forme non aléatoire, alors le modèle n'est pas linéaire, les erreurs ne sont pas normalement distribuées, n'ont pas une variance constante (hétérocédastique) et sont autos corrélées

Inférence sur β_1

b_1 est un estimateur efficace de β_1, b_1 étant un estimateur, il possède une distribution d'échantillonnage avec une moyenne et une variance

Distribution d'échantillonnage de b_1

si $\varepsilon \in N(0, {\sigma_E}^2)$ et $Cov(e_i, e_j) = 0$ et $Cov(e_i, x_i) = 0$, alors $b_1 \in N\left(\beta_1, {\sigma_E}^2\left(\frac{1}{\sum(X_i-\bar{X})^2}\right)\right)$

Estimation de la variance $\sigma^2(b_1)$

b_1 distribué normalement, taille d'échantillon petite, $\sigma^2(b_1)$ inconnue

$S^2(b_1) = \frac{{S_e}^2}{\sum(X_i-\bar{X})^2}$, $S(b_1) = \frac{S_e}{\sqrt{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}}}$ et $t = \frac{b_1-\beta_1}{s(b_1)} \in t(n-2)$

Intervalle de confiance pour β_1

$b_1 - t_{\alpha/2;n-2}s(b_1) \le \beta_1 \le b_1 + t_{\alpha/2;n-2}s(b_1)$, $S(b_1) = \frac{S_e}{\sqrt{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}}}$

Test d'hypothèse pour β_1

Le paramètre β_1 mesure la dépendance linéaire entre Y et X
Si cette dépendance est significative, alors $\beta_1 \neq 0$
Si $\beta_1 = 0$, le modèle devient : $Y_i = \beta_0 + \varepsilon_i$ ou $Y_i = E(Y_i) + \varepsilon_i$

- $H_0 : \beta_1 = 0$ $\quad H_1 : \beta_1 \neq 0$
- Conditions du test : le modèle est valide, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, b_1 distribué selon la loi normale, ${\sigma^2}(b_1)$ inconnue, α fixé
- La statistique : $t = \frac{b_1-\beta_1}{s(b_1)} = \frac{b_1}{s(b_1)} \in t(n-2)$
- La règle de décision : i) $H_1 : \beta_1 \neq 0$, Rejeter H_0 si $t < -t_{\alpha/2;n-2}$ ou si $t > t_{\alpha/2;n-2}$

Inférence sur β_0

b_0 est un estimateur efficace de β_0, b_0 étant un estimateur, il possède une distribution d'échantillonnage avec une moyenne et une variance

Distribution d'échantillonnage de b_0

si $\varepsilon \in N(0, {\sigma_E}^2)$ et $Cov(e_i, e_j) = 0$ et $Cov(e_i, x_i) = 0$, alors $b_0 \in N\left(\beta_0, {\sigma_E}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right)\right)$

Estimation de la variance $\sigma^2(b_0)$

$S^2(b_0) = {S_e}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right)$,S= $S_e\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}}}$et $t = \frac{b_0-\beta_0}{s(b_0)} \in t(n-2)$

Intervalle de confiance pour β_0

- $b_0 - t_{\alpha/2;n-2}s(b_0) \le \beta_0 \le b_0 + t_{\alpha/2;n-2}s(b_0)$
- $S(b_0) = S_e\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}}}$

Test d'hypothèse pour β_0

Si le test d'hypothèse sur β_0 favorise l'hypothèse nulle H_0: $\beta_0 = 0$, le modèle devient : $Y_i = \beta_1 X_i + \varepsilon_i$
Dans ce cas, la droite de régression passe par l'origine

- $H_0 : \beta_0 = 0$ $\quad H_1 : \beta_0 \neq 0$
- Conditions du test : le modèle est valide, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, b_0 distribué selon la loi normale, ${\sigma^2}(b_0)$ inconnue, α fixé
- La statistique : $t = \frac{b_0}{s(b_0)} \in t(n-2)$
- La règle de décision : i) $H_1 : \beta_0 \neq 0$, Rejeter H_0 si $t < -t_{\alpha/2;n-2}$ ou si $t > t_{\alpha/2;n-2}$

Intervalle de confiance pour $E(Y_h)$

- Pour $X = X_h$, l'estimation ponctuelle de $E(Y_h)$ est $\hat{Y}_h = b_0 + b_1 X_h$
- La distribution d'échantillonnage de $E(Y_h)$ est $\frac{\hat{Y}_h - E(Y_h)}{S(\hat{Y}_h)} \in t(n-2)$
- L'intervalle de conf : $\hat{Y}_h - t_{\alpha/2;n-2}\,S(\hat{Y}_h) \le E(Y_h) \le \hat{Y}_h + t_{\alpha/2;n-2}\,S(\hat{Y}_h)$

Calcul de $S(\hat{Y}_h)$

$S^2(\hat{Y}_h) = S^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$, $S(\hat{Y}_h) = S\sqrt{\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}} = S\sqrt{\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}}}$

Intervalle de confiance pour Y_h

Pour $X = X_h$, l'estimation ponctuelle de Y_h est $\hat{Y}_h = b_0 + b_1 X_h$
La distribution d'échantillonnage de Y_h es $t = \frac{Y_h - \hat{Y}_h}{S\sqrt{1 + \frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}}} \in t(n-2)$

L'intervalle de conf : $\hat{Y}_h - t_{\alpha/2;n-2}\,S(d_h) \le Y_h \le \hat{Y}_h + t_{\alpha/2;n-2}\,S(d_h)$

Calcul de $S(d_h)$

$S^2(d_h) = S^2 + \frac{S^2}{n} + S^2\frac{(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2}$, $S(d_h) = S\sqrt{1 + \frac{1}{n} + \frac{(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2}} =$

$S\sqrt{1 + \frac{1}{n} + \frac{(X_i-\bar{X})^2}{\sum {x_{v.2}}_i - \frac{(\sum X_i)^2}{n}}}$

EXCEL

Chapitre 8

Test d'égalité des espérances pour 2 pop de variances égales		
CAS 3 – test de 2 moyenne	**Classe 1**	**Classe 2**
Moyenne	$\bar{X}_1$	$\bar{X}_2$
Variance	${S_1}^2$	${S_2}^2$
Observations	n_1	n_2
Variance pondérée	${S_c}^2 = \text{Var}(\bar{X}_1 - \bar{X}_2)$	
Différence hypothétique…		-
Degré de liberté	n-2	-
Statistique t	Résultat empi.	-
P(T<=t) unilatéral		-
Valeur critique de t (unilatéral)	t calculé	-
P(T<=t) bilatéral		-
Valeur critique de t (bilatéral)		-
Test d'égalité des espérances : observations pairées		
Moyenne	$\bar{X}_1$	$\bar{X}_2$
Variance	${S_1}^2$	${S_2}^2$
Observations	n_1	n_2
Coefficient de cor. de Pearson		-
Différence hypothétique…		-
Degré de liberté	n-1 *(1 n slmt)*	-
Statistique t	Résultat empi.	-
P(T<=t) unilatéral		-
Valeur critique de t (unilatéral)	t calculé	-
P(T<=t) bilatéral		-
Valeur critique de t (bilatéral)		-

Moyenne	$\bar{x}$
Erreur-type	S_e
Écart-type	σ
Variance de l'échantillon	σ^2
Plage	Max - min
Min	# min par observation
Max	# max par observation
Somme	Total d'échantillon
Nombre d'observations	n
Test d'égalité des variances (F-test)	
Moyenne	$\bar{x}_1$ $\quad$ $\bar{x}_2$
Variance	${S_1}^2$ $\quad$ ${S_2}^2$
Observation	n_1 $\quad$ n_2
Degré de liberté	$n_1 - 1$ $\quad$ $n_2 - 1$
F	Résultat empi. $({S_1}^2/{S_2}^2)$
P(F<=f) unilatéral	-
Valeur critique F (unilatéral)	t calculé -

Chapitre 9

Résultat empirique : calcul de χ^2						
	$F_{obs.}$	$F_{théor.}$		Calcul de χ^2		
Résultats	f_{oi}	P_i	f_{ti}	f_{oi}-f_{ti}	$(f_{oi}-f_{ti})^2$	$(f_{oi}-f_{ti})^2/f_{ti}$
1						
2						
Total						$\chi^2 =$

Chapitre 10

Analyse de variance pour un facteur

RAPPORT DETAILLE

Groupe	# échantillon	Somme	Moyenne	Variance
1	n_1		$\bar{x}_1$	${S_1}^2$
2	n_2		$\bar{x}_2$	${S_2}^2$
3	n_3		$\bar{x}_3$	${S_3}^2$

ANALYSE DE VARIANCE

Source des variations	$\sum$ des carrés	*dl*	$\bar{x}$ des carrés	F	Prob	Valeur crit F
Entre Groupe	SC_A *	k-1 *2	CM_A = *÷*2	$\frac{CM_A}{CM_{RES}}$	(P-val.)	F calculé
À l'intérieur des groupes	SC_{RES} *1	n-k *3	CM_{RES} = *1÷*3	-	-	-
Total	SCT	n-1	-	-	-	-

Chapitre 11

Calcul préliminaires

j	y_i	x_i	x_iy_i	${y_i}^2$	${x_i}^2$
1					
2					
3					
Total					

Analyse de variance : régression linéaire simple

RAPPORT DETAILLE

Statistiques de la régression	
Coefficient de détermination multiple	r
Coefficient de détermination R^2	r^2
Coefficient de détermination R^2	
Erreur-type	S_e
Observations	n

	Coefficients	Erreur-type	Statistique t	Probabilité	Limite inf 95%	Limite Sup. 95%
Constante	b_0	b_0, S(b_0)	Résultat empirique	**plus petite que 5% = significatif**		
Variable X1	b_1	b_1, S(b_1)	Résultat empirique			

$CM_{RÉS}$ est toujours égale à ${S_e}^2$,

ANALYSE DE VARIANCE

	Degrés de liberté	$\sum$ des carrés	$\bar{x}$ des carrés	F	Valeur critique pour F (P-value)
Régression	1	*SCR*	*divisé/dl*	<- / <-l	F calculé →avec cette valeur on peut savoir si on rejette H_0 (5%)
Résidus	n-2	SC_{RES}	*divisé/dl*	-	-
Total	n-1	total	-	-	-

CHAPITRE 11

- Si 2 variables aléatoires sont statistiquement indépendantes (aucune liaison entre elles), le coefficient de corrélation est nul (Toutefois la réciproque n'est pas nécessairement vraie)
- L'hypothèse de normalité des erreurs n'est pas nécessairement pour estimer les paramètres β_0 et β_1 par la méthode des moindres²
- Si toutes les observations sont sur la droite de régression alors les résidus sont nuls, $SC_{rés} = 0$, $SCT = SCR$ *et* $r^2 = 1$
- Si aucune relation existe entre Y et X, alors SCR=0 et $r^2 = 0$, **DONC** $\mathbf{0 \le r^2 \le 1}$

CHAPITRE 8 – Test d'hypothèse sur deux paramètres

Différence de deux moyennes

- Soit X_1 et X_2, deux variables aléatoires de moyenne μ_1 *et* μ_2 et de variance ${\sigma_1}^2$ *et* ${\sigma_2}^2$ respectivement
- La statistique $(\bar{X}_1 - \bar{X}_2)$ est utilisée pour comparer l'espérance mathématique $(\mu_1 - \mu_2)$ des deux populations

Cas 1 : Pop. normales, et ${\sigma_1}^2$ *et* ${\sigma_2}^2$ connues

Si on prélève indépendamment des échantillons aléatoires de taille n_1 et n_2 respectivement de 2 pop normales tel que $X_1 \in N(\mu_1, {\sigma_1}^2)$ *et* $X_2 \in N(\mu_2, {\sigma_2}^2)$ alors la distribution d'échantillonnage de l'estimateur $(\bar{X}_1 - \bar{X}_2)$ est une loi nor. avec :

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \text{ et } Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Soit X_1 et X_2, 2 V.A. indépendantes, tel que $X_1 \in N(\mu_1, {\sigma_1}^2)$ *et* $X_2 \in N(\mu_2, {\sigma_2}^2)$ alors : $(\bar{X}_1 - \bar{X}_2) \in N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ et

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in N(0,1)$$

Cas 2 : Grands échantillons, $n_1 \geq 30$ et $n_2 \geq 30$

Si on prélève indépendamment des échantillons aléatoires de taille n_1 et n_2 suffisamment grandes, alors grâce au théorème central limite, la différence $(\bar{X}_1 - \bar{X}_2)$ suit approximativement une loi normale avec :

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \text{ et } Var(\bar{X}_1 - \bar{X}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

Si X_1 et X_2, sont 2 V.A ind. et que n_1 et n_2 sont grands, alors on a approx. $(\bar{X}_1 - \bar{X}_2) \in N\left(\mu_1 - \mu_2, \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)$ et $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \in N(0,1)$

Cas 3 : Pop. normales, ${\sigma_1}^2$ *et* ${\sigma_2}^2$ inconnues mais supposées =, et n_1 et/ou n_2 petits (si les var sont ≠, faut utiliser le test de Satterhwaite, cas 4)

Si on prélève indépendamment des échantillons aléatoires de petites tailles (n_1 et/ou $n_2 < 30$) de populations normales de variances inconnues mais supposées égales à σ^2, on obtient :

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \text{ et } Var(\bar{X}_1 - \bar{X}_2) = \frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Si $X_1 \in N(\mu_1, {\sigma_1}^2)$ *et* $X_2 \in N(\mu_2, {\sigma_2}^2)$, ${\sigma_1}^2$ *et* ${\sigma_2}^2$ sont inconnues mais supposées égales à σ^2, et n_1 et/ou n_2 sont petits, alors on a :

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \in t(n_1+n_2-2)$$

Estimation de σ^2 par ${S_c}^2$ & Calcul de $Var(\bar{X}_1 - \bar{X}_2)$

$${S_c}^2 = \frac{\sum(X_{i1}-\bar{X}_1)^2 + \sum(X_{i2}-\bar{X}_2)^2}{(n_1-1)+(n_2-1)} = \frac{\sum(X_{i1}-\bar{X}_1)^2 + \sum(X_{i2}-\bar{X}_2)^2}{(n_1+n_2-2)} = \frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{(n_1+n_2-2)}$$

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{(n_1+n_2-2)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{{S_c}^2}{n_1} + \frac{{S_c}^2}{n_2} = {S_c}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Test sur une différence de moyennes $(\mu_1 - \mu_2)$

- $\mathbf{H_0}: \mu_1 - \mu_2 = 0$, $\mathbf{H_1}: \mu_1 - \mu_2 \neq 0; \mu_1 - \mu_2 < 0; \mu_1 - \mu_2 > 0, \alpha$ fixé
- Conditions du test :
  - **CAS 1**; 2 E.A. ind., pop. Norm. et ${\sigma_1}^2$ *et* ${\sigma_2}^2$ connues
  - **CAS 2**; 2 E.A. ind. de grande taille **(CHANGE σ POUR S)**
- La statistique pour : $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in N(0,1)$
- La règle de décision : i) $H_1: \mu_1 - \mu_2 \neq 0$, Rejeter H_0 si $(\bar{X}_1 - \bar{X}_2) < (\mu_1 - \mu_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ ou si $(\bar{X}_1 - \bar{X}_2) > (\mu_1 - \mu_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

  ii) $H_1: \mu_1 - \mu_2 < 0$, Rejeter H_0 si $(\bar{X}_1 - \bar{X}_2) < (\mu_1 - \mu_2) - z_\alpha\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

  iii) $H_1: \mu_1 - \mu_2 > 0$, Rejeter H_0 si $(\bar{X}_1 - \bar{X}_2) > (\mu_1 - \mu_2) + z_\alpha\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

CAS 3; 2 E.A. ind., ${\sigma_1}^2$ *et* ${\sigma_2}^2$ inconnues, et n_1 et/ou n_2 petits

- La statistique pour : $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \in t(n_1+n_2-2)$
- La règle de décision : i) $H_1: \mu_1 - \mu_2 \neq 0$, Rejeter H_0 si $(\bar{X}_1 - \bar{X}_2) < (\mu_1 - \mu_2) - t_{\alpha/2}\sqrt{\frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ ou si $(\bar{X}_1 - \bar{X}_2) > (\mu_1 - \mu_2) + t_{\alpha/2}\sqrt{\frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- La règle de décision : ii) $H_1: \mu_1 - \mu_2 < 0$

  Rejeter H_0 si $(\bar{X}_1 - \bar{X}_2) < (\mu_1 - \mu_2) - t_\alpha\sqrt{\frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- La règle de décision : iii) $H_1: \mu_1 - \mu_2 > 0$

  Rejeter H_0 si $(\bar{X}_1 - \bar{X}_2) > (\mu_1 - \mu_2) + t_\alpha\sqrt{\frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Test d'hypothèse pour observations couplées

- On veut comparer les moyennes des 2 caractéristiques provenant de $\hat{m}$ unités stats d'une pop. --- Les variables X_1 et X_2 ne sont plus indépendantes --- Les observations sont couplées

Observations couplées

- Soit les observations pour $X_1 : X_{1,1}, X_{2,1}, \dots, X_{n,1}$
- Soit les observations pour $X_2 : X_{1,2}, X_{2,2}, \dots, X_{n,2}$
- Ces observations sont couplées : $(X_{1,1}, X_{1,2}), (X_{2,1}, X_{2,2}), (X_{n,1}, X_{n,2})$
- On calc la dif pour chak couple $i : D_i = X_{i,1} - X_{i,2}$ pour $i = 1, 2, \dots, n$
- Si D_i provient d'une pop. d'espérance μ_D et de variance ${\sigma_D}^2$ alors : $\bar{D} = \frac{\sum D_i}{n}$ est un estim de μ_D $\quad {S_D}^2 = \frac{\sum(D_i - \bar{D})^2}{n-1}$ est un estim de ${\sigma_D}^2$

  $\bar{D}$ est la valeur empirique

Inférence pour observations couplées

- Cas 1 : pop. normale, et ${\sigma_D}^2$ connue
- Cas 2 : grand échantillon, $n \geq 30$
- Cas 3 : pop. normale, ${\sigma_D}^2$ inconnue, et n est petit

Test pour μ_D

$H_0: \mu_D = 0 \qquad H_1: \mu_D \neq 0;\ \mu_D < 0;\ \mu_D > 0 \qquad \alpha$ fixé

- Conditions du test : 2 É.A. avec observ. couplées **[cas 1, 2 ou 3]**
- La statistique **[cas 1, cas 2** = $\sigma \to S$, **cas 3** = $\sigma \to S, Z \to t$] :

  $Z = \frac{\bar{D} - \mu_D}{\sigma_D} \in N(0,1)$ [cas 3 : $\in t(n-1)$ù]
- La règle de décision : i) $H_1: \mu_D \neq 0$

  Rejeter H_0 si $\bar{D} < \mu_D - z_{\alpha/2}\sqrt{\frac{\sigma_D^2}{n}}$ ou si $\bar{D} > \mu_D + z_{\alpha/2}\sqrt{\frac{\sigma_D^2}{n}}$ ou=$\left(\frac{\sigma_D}{\sqrt{n}}\right)$
- La règle de déc : ii) $H_1: \mu_D < 0$; Rejeter H_0 si $\bar{D} < \mu_D - z_\alpha\sqrt{\frac{\sigma_D^2}{n}}$
- La règle de déc : iii) $H_1: \mu_D > 0$; Rejeter H_0 si $\bar{D} > \mu_D + z_\alpha\sqrt{\frac{\sigma_D^2}{n}}$

Différence entre deux proportions

- Soit X_1 et X_2 2 V.A. ind distribuées selon une loi binomiale de paramètre p_1 et p_2 respectivement et provenant de 2 pop. dif.
- On veut comparer la proportion de succès $(p_1 - p_2)$ entre les 2 pop

Soit des EA ind taille n_1 *et* n_2 provenant 2 pop bin. de p_1 *et* p_2 alors : $(\hat{P}_1 - \hat{P}_2) \in N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$, autant n_1 *et* n_2 sont grands

Donc en autant que n_1 *et* n_2 sont ↑, on a : $E\left(\hat{P}_1 - \hat{P}_2 \to valeur\ emp.\right) = p_1 - p_2 \qquad Var\left(\hat{P}_1 - \hat{P}_2\right) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

Test d'hypothèse pour $(p_1 - p_2)$ - α fixé (p est un %)

$\mathbf{H_0}: p_1 - p_2 = 0\ (p_1 = p_2 = p)$ $\mathbf{H_1}: p_1 - p_2 \neq 0; p_1 - p_2 < 0; p_1 - p_2 > 0$

Conditions du test : deux échantillons aléatoires indépendants, n_1 *et* n_2 grands ($n_1p_1, n_2p_2, n_1(1-p_1), n_2(1-p_2)$*soient tous* ≥ 5)

Puisque $p_1 = p_2 = p$, on doit estimer p : $\hat{P} = \frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1 + n_2}$

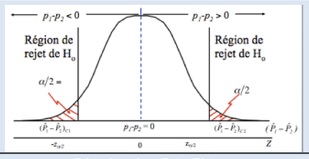
La statistique à utiliser est $Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \in N(0,1)$

Puisque selon H_0 on $p_1 = p_2 = p$, stat devin $Z = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \in N(0,1)$

La règle de décision : i) $H_1: p_1 - p_2 \neq 0$, Rejeter H_0 si $\left(\hat{P}_1 - \hat{P}_2\right) < -z_{\alpha/2}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ ou si $\left(\hat{P}_1 - \hat{P}_2\right) > z_{\alpha/2}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

ii) $H_1: p_1 - p_2 < 0$, Rejeter H_0 si $\left(\hat{P}_1 - \hat{P}_2\right) < -z_\alpha\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

iii) $H_1: p_1 - p_2 > 0$, Rejeter H_0 si $\left(\hat{P}_1 - \hat{P}_2\right) > z_\alpha\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$



Distribution F de Fisher

Pour utiliser loi de Fisher, les obs doivent provenir de lois normales

La loi de Fisher est formée par le rapport entre 2 V.A. indépendantes suivant une distribution de Khi-deux divisées par leurs *dl* $0 \leq F \leq \infty$

Propriétés de loi Fisher

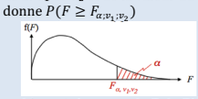
La courber de la loi de Fisher a une asymétrie positive

La distribution de Fisher dépend uniquement de v_1 *et* v_2

Table du F de Fisher

La table de la loi de Fisher donne les probabilités, α, en fonction du nombre de *dl* v_1 *et* v_2 et de la valeur du F

La table donne $P(F \geq F_{\alpha;v_1;v_2})$



Relation complémentaire pour le F de Fisher – si on ne peut pas l'utiliser on se sert de $F_{1-\alpha;v_1;v_2} = \frac{1}{F_{\alpha;v_1;v_2}}$

Distribution du quotient de deux variances ${S_1}^2/{S_2}^2$

On veut comparer la variance de deux populations

On tire un échantillon de taille n_1 *et* n_2 respectivement de 2 pop norm

Pour comparer leur variance on se sert du quotient ${S_1}^2/{S_2}^2$

Si la variable analysée pour les deux pop suit une distribution de probabilité normale alors : $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \in F(n_1-1, n_2-1), 0 \leq F \leq \infty$

Test d'hypothèse pour deux variances

$\mathbf{H_0}: \frac{\sigma_1^2}{\sigma_2^2} = 1$ -- $\mathbf{H_1}: \sigma_1^2/\sigma_2^2 \neq 1; \sigma_1^2/\sigma_2^2 < 1; \sigma_1^2/\sigma_2^2 > 1$ -- α fixé

Conditions du test : 2 échantillons aléatoires indépendantes, pop norm

La stat : $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \in F(n_1-1, n_2-1)$ Sous H_0, $F = \frac{S_1^2}{S_2^2} \in F(n_1-1, n_2-1)$

La règle de décision : i) $H_1: \sigma_1^2/\sigma_2^2 \neq 1$, Rejeter H_0 si $\frac{S_1^2}{S_2^2} < \frac{1}{F_{\alpha/2}(n_2-1, n_1-1)}$ ou si $\frac{S_1^2}{S_2^2} > \frac{1}{F_{\alpha/2}(n_1-1, n_2-1)}$

ii) $H_1: \sigma_1^2/\sigma_2^2 < 1$, Rejeter H_0 si $\frac{S_1^2}{S_2^2} < \frac{1}{F_\alpha(n_2-1, n_1-1)}$

iii) $H_1: \sigma_1^2/\sigma_2^2 > 1$, Rejeter H_0 si $\frac{S_1^2}{S_2^2} > \frac{1}{\quad}$

Chapitre 9 – Khi-deux

Tableau croisé et test d'indépendance

1. On veut tester l'hypothèse selon laquelle deux variables X et Y mesurées sur une échelle nominale ou ordinale sont ind
2. Les obs sont réparties selon les modalités croisées des 2variables
3. On obtient un tableau à 2dimensions appelé tableau croisé ou tableau de contingence

Structure de tableau croisé

	Y				Total de la ligne
X	B_1	B_2	...	B_k	
A_1	f_{o11}	f_{o12}	...	f_{o1k}	L_1
A_2	f_{o21}	f_{o22}	...	f_{o2k}	L_2
...	...	...	...	...	...
A_r	f_{or1}	f_{or2}	...	f_{ork}	L_r
Total de la colonne	C_1	C_2	...	C_k	n

Les fréquences observées

1. $f_{o_{ij}}$: fréquence observée des données ayant la modalité A_i de la variable X et la modalité B_j de la variable Y
2. $C_j = \sum_{i=1}^{r} f_{o_{ij}}, j = 1,2, \dots, k$
3. $L_i = \sum_{j=1}^{k} f_{o_{ij}}, i = 1,2, \dots, r$
4. $\sum_{i=1}^{r} L_i = \sum_{j=1}^{k} C_j = \sum_{i=1}^{r}\sum_{j=1}^{k} f_{o_{ij}} = n$
1. H_0 : les deux caractères X et Y sont indépendants
2. H_1 : les deux caractères X et Y ne sont pas indépendants
3. Conditions du test : Seuil de signification du test, α, est fixé, Échantillon aléatoire de taille n, Les fréquences théoriques sont supérieures ou égales à 5 (les fréquences observées sont tjrs données)

   Calcul du χ^2 pour le test d'indépendance (calcul chak donnée)

   $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(f_{o_{ij}} - f_{t_{ij}})^2}{f_{t_{ij}}} \in \chi^2((r-1)(k-1))$ où $f_{t_{ij}} = \frac{L_i \times C_j}{n}$
1. On rejette H_0 si $\chi^2 > \chi^2_{\alpha;(r-1)(k-1)}$ (χ^2 théorique > χ^2 calculée)

Mesure d'association entre deux caractères

1. Coefficient de contingence v de Cramer, $v = \sqrt{\frac{\chi^2}{n(k-1)}}$, où k est le min entre le # de lignes et le nombre de colonnes et $0 \leq v \leq 1$

Rappel : P(A∩B) = P(A) x P(B) seulement si indépendant

Test d'ajustement analytique (test du χ^2 de Pearson)

2. On veut vérifier si une distribution de fréquence empirique provient d'une distribution théorique connue
3. Le test du χ^2 permet de vérifier s'il y a une différence significative entre les f_o et les f_{calc} selon la distribution théorique

Distribution de fréquences observées

4. Les observations de l'échantillon sont réparties en classe (distribution de fréquences)
   1. $LI_1 \leq X \leq LS_1, F_{o_1}$
   2. $LI_2 \leq X \leq LS_2, F_{o_2}$ ...
   3. $LI_k \leq X \leq LS_k, F_{o_k}$ → fréquence observée dans la classe k
   4. $\sum_{i=1}^{k} F_{o_1} = n$, où n : taille de l'échantillon

Distribution de fréquences théoriques

5. Calculer les fréquences théoriques selon H_0
   1. $p_1 = P(LI_1 \leq X \leq LS_1), F_{t_1} = p_1 n$
   2. $p_2 = P(LI_2 \leq X \leq LS_2), F_{t_2} = p_2 n$ ...
   3. $p_k = P(LI_k \leq X \leq LS_k), F_{t_k} = p_k n$
   4. $\sum_{i=1}^{k} p_i = 1, \sum_{i=1}^{k} F_{t_i} = n$

Test d'ajustement analytique

6. H_0 : les obsv proviennent de la distribution théorique $f(x;\theta)$
   7. lois discrètes (binomiale, Poisson, ...)
   8. lois continues (exponentielle, uniforme, normale, ...)
9. H_1 : les obsv ne proviennent pas de la distribution théorique $f(x;\theta)$
10. α : le seuil de signification est fixé
11. Conditions du test (même que le test d'indépendance + Échantillon aléatoire, n est grand ($n \geq 50$), tel que les $F_{t_i} \geq 5$, (on peut regrouper les classes aux extrémités de la distribution pour respecter cette condition)

1. Le statistique : $\chi^2 = \sum_{i=1}^{k} \frac{(f_{o_i} - F_{t_i})^2}{F_{t_i}} \in \chi^2(k-1-r)$, où r est le # de paramètres à estimer de la distribution théoriqu et k le # de classes
2. Région critique : rejeter H0 si $\chi^2 > \chi^2_{\alpha;k-1-r}$
3. Résultat empirique : calculer $\chi^2 = \sum_{i=1}^{k} \frac{(f_{o_i} - F_{t_i})^2}{F_{t_i}}$

Le rejet de H_o peut provenir de trois causes:

1. la durée de vie ne suit pas une loi normale
2. la moyenne de la durée de vie n'est pas m
3. l'écart-type de la durée de vie n'est pas s

Les causes 2) et 3) peuvent se vérifier par les tests appropriés sur une moyenne ou sur une variance

Distribution binomiale - # de succès dans l'échantillon

$X \in b(x;n;p) = C_x^n p^x (1-p)^{n-x}$, pour $x = 0, 1, 2, \dots$

$\bar{x} = \frac{\sum_{i=1}^{k} f_{o_i} x_i}{\sum_{i=1}^{k} f_{o_i}}, \quad \hat{p} = \frac{\bar{x}}{n}$

Distribution de Poisson

$X \in P(x;\lambda)$

$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, \dots$

$\bar{\lambda} = \frac{\sum_{i=1}^{k} f_{o_i} x_i}{\sum_{i=1}^{k} f_{o_i}}$

Chapitre 10 – Analyse de variance (ADV)

Comparaison de plusieurs moyennes

- On veut comparer plus de deux moyennes
- Le *t* de Student ne peut plus être utilisé
- L'ADV (ANOVA) permet de tester l'= des moyennes de 2 ou + pop
- $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

  H_1 : les μ_j ne sont pas toutes égales $j = 1, \dots, k$

Schématisation des données, où

- $\bar{y}$ est la moyenne de toutes les observations
- $\bar{y}_1$ est la moyenne de toutes les observations dans le groupe 1
- **Rou** = la VAR tot; **JAUNE** = VAR ds le group ; **BLEU** = VAR entre groupes

$n = n_1 + n_2 + \dots + n_k$

	Modalité				
	1	**2**	...	***k***	
résultats	y_{11}	y_{12}	...	y_{1k}	
	y_{21}	y_{22}	...	y_{2k}	
	...	...		...	
	$y_{n_1 1}$	$y_{n_2 2}$	...	$y_{n_k k}$	
	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_k$	$\bar{y}$

Décomposition de la variation, où $\bar{y}$ est la moyenne globale, y_{ij} une donnée (moyenne individuelle) et $\bar{y}_j$ la moyenne d'un groupe

- Écart total = écart attribuable au facteur + écart résiduel
- $(y_{ij} - \bar{y}) = (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)$ (& au carré)

Détermination des sommes de carrés

- $(y_{ij} - \bar{y})^2 = (\bar{y}_j - \bar{y})^2 + (y_{ij} - \bar{y}_j)^2$

  $= \sum_{j=1}^{n_j}\sum_{i=1}^{n_j}(\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$
- $SCT = SC_A$ (*entre les groupes*) + SC_{RES} (*à l'intérieur d'un groupe*)

Expression des sommes de carrés

Somme de carrés totale, $SCT = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$

$= \sum_{j=1}^{k}\sum_{i=1}^{n_j} {y_{ij}}^2 - \frac{T^2}{n}$, où $\boldsymbol{T} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \boldsymbol{y_{ij}}$

Somme de carrés attribuable au facteur A, $SC_A = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{k}\frac{T_j^2}{n_j} - \frac{T^2}{n}$, où $\boldsymbol{T_j} = \sum_{i=1}^{n_j} \boldsymbol{y_{ij}}$

Somme de carrés résiduelle, $SC_{RES} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} {y_{ij}}^2 - \sum_{j=1}^{k}\frac{T_j^2}{n_j}$

Calcul des degrés de liberté

- $SCT = SC_A + SC_{RES}$, somme de carrés
- $(n-1) = (k-1) + (n-k)$, degrés de liberté
  - n : taille de l'ensemble des échantillons (ensemble x échantillon)
  - k : nombre d'échantillons

Calcul des carrés moyens

Carré moyen dû au facteur A, $CM_A = \frac{SC_A}{k-1} = \frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2}{k-1}$

Carré moyen résiduel, $CM_{RES} = \frac{SC_{RES}}{n-k} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{n-k}$

Table d'analyse de variance

Source de variation	Somme de carrés	Degrés de liberté	Carrés moyens	Rapport *F*
Attribuable au facteur A (expliquée)	SC_A	$k-1$	CM_A	
Résiduelle (inexpliquée)	SC_{RES}	$n-k$	CM_{RES}	
Totale	SCT	$n-1$		

TESTS

- $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

  H_1 : les μ_j ne sont pas toutes égales $j = 1, \dots, k$
- **Alpha** : le seuil de significatif est fixé
- Condition d'application : échantillons aléatoires provenant de k pop norm. de VAR identique σ^2
- La statistique : $F = \frac{CM_A}{CM_{RES}}$
- Règle de décision : rejeter H0 si $F > F_{\alpha;(k-1),(n-k)}$

Méthode de Tukey-Kramer

- Ordonner par valeurs crois. les k moyennes obtenues pour les k pop
- Utiliser les carré moyen résiduel et ses *dl* (CM_{RES} *avec* $(n-k)$*dl*)
- Utiliser la table Q de Tukey-Kramer (prend valeur la plus près de v)
- Calculer l'écart critique $w = Q_{0,05;k,(n-k)}\sqrt{\frac{CM_{RES}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$
- Si $n_1 = n_2 = \dots = n_k$, alors $w = Q_{0,05;k,(n-k)}\sqrt{\frac{CM_{RES}}{n_k}}$,
- Comparaison des écarts entre les moyennes des modalités deux à 2

Chapitre 12 – Régression linéaire multiple

Modèle de régression linéaire multiple

- $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$
  - Y_i : variable dépendante
  - X_{ij} : variable indépendante j
  - ε_i : composante aléatoire
  - $\beta_0, \beta_1, \dots, \beta_k$: paramètres à estimer

Hypothèses fondamentales

- ε_i est une V.A. dont $E(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \sigma_\varepsilon^2$ pour tout X_i
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ pour tout i,j et $i \neq j$.
- $X_1, X_2, \dots, X_k$ non aléatoires et non corrélées
- ε_i suit une distribution normale, $N(0, \sigma_\varepsilon^2)$
- Restriction : n > k + 1

La droite de régression

- $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}$

Problème de multi colinéarité

- Lorsqu'il y a plusieurs variables indépendantes, il arrive souvent que ces variables soient corrélées entre elles (multi colinéarité).
- La multi colinéarité des VAR indépenda. occasionne des problèmes pour estimer correctement les paramètres du modèle de régression.
- Peut alors introduire VAR l'une après autre et choisir plus significa.
- L'analyse factorielle p-e utilisée pour identifier facteurs principaux

LÉGENDE/FORMULES

- **Loi normale = intervalle jusqu'à l'infini**
- **Loi de Student : $t_{\alpha;n-1}$**
- **N = taille au total, n = unités statistiques (échantillon que l'on veut)**
- **k = fréquence de sélection = N/n = # de modalité**
- **S_e = erreur type**
- ***dl* = degré de liberté, *v* (*n*-1)**
- **α = obtenir aléatoirement**
- **θ = paramètre à estimer**
- **σ^2 = variance de (x) = $s^2 = \frac{\sum(x_i - \mu)^2}{N}$**
- **σ = écart-type = s**
- **1 – α (= niveau de confiance) $P(LI \leq \theta \leq LS)$**
- **α = seuil de signification**
- **μ = E(X) = $\frac{\sum x_i}{N}$ ($E(\bar{X}) = \mu$) = moyenne observée**
- **$Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\sigma^2}{n} = \frac{(\bar{x} - E(\bar{X}))^2 \times f}{Fréquence\ (totale)}$**
- **Toujours rejeter H_0 si $\alpha_p < \alpha$**
- **SCR = variation expliquée**
- **SC_{res} = Variation inexpliquée**

Élaboration d'un test d'hypothèse

1. **Énoncer H_0 et H_1**
2. **Choisir le seuil de signification *a***
3. **Énoncer les conditions du test (forme de la distribution, variance connue ou inconnue, taille de l'échantillon, ...)**
4. **Déterminer la statistique compatible avec H_0 et l'écart réduit**
5. **Déterminer la règle de décision (la région critique)**
6. **Calculer l'écart réduit empirique**
7. **Conclure**