2V481 – BM2

Cours 2

LICENCE SCIENCES ET TECHNOLOGIE
MENTION SCIENCE DE LA VIE – L2

LARSEN MARTIN
DEMEYRIER VIRGINIE
RYBARCZYK HERVÉ

One-way ANOVA
ou
Analyse de variances

LICENCE SCIENCES ET TECHNOLOGIE
MENTION SCIENCE DE LA VIE – L2

Choix de test dépend de la caractéristique de l'étude

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	X^2 or bi-nominal	X^2	Macnarmar's X^2	X^2	Cochran's Q	
Rank or Ordinal	X^2	Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Parametric (Interval & Ratio)	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r
		Factorial (2 way) ANOVA				

3

Analyse de variances

Test paramétrique de **Comparaison de moyennes** de a échantillons indépendants

Exemple : 4 traitements différents => 1 variable qualitative à 4 niveaux/modalités.

Pour chaque niveau, on réalise un certain nombre d'observations sur la **variable dépendante**.
Exemple : poids des individus après le traitement.

Question associée : Y-t-il une différence entre les moyennes obtenues en fonction du traitement ?

Dès que plus de 2 traitements, les tests de comparaison de moyennes ne sont plus appropriés.
⇒ On compare globalement toutes les moyennes de a échantillons en 1 seul test.
⇒ C'est l'objectif de l'analyse de variance ou **ANOVA** (pour les petits et les grands échantillons).
⇒ On utilise les **variances** pour mettre en évidence les différences entre les moyennes.
⇒ Détermine si 1 facteur ou groupe a un effet sur la variable quantitative.

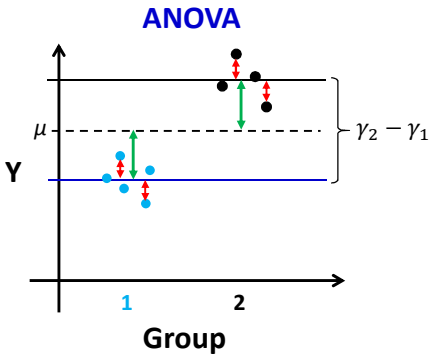
4

Analyse de variances

⇒ Analyse les sources de variation de l'ensemble des données (dispersion des moyennes due ou non aux fluctuations d'échantillonnage).

VA Y se décomposant en : **Y = μ + γ_i + e_{ij}**

Avec :
μ : l'espérance de la population d'où est extrait l'échantillon sur lequel on mesure Y.
γ_i : l'effet du facteur, s'il existe, modifiant la valeur de la variable selon l'intensité du facteur appliqué.
e_{ij} : le résidu, c'est-à-dire la variation individuelle aléatoire.



L'ANOVA repose sur une décomposition de la variabilité (ou dispersion) en :

$$\boldsymbol{Y} = \mu + \begin{pmatrix} Treat & Coeff \\ 1 & \gamma_1 \\ 2 & \gamma_2 \end{pmatrix} + \varepsilon$$

Variabilité Totale = Variabilité due au facteur étudiée + Variabilité restante ou résiduelle

5

Analyse de variances

- Conditions d'application :
  - Tirage aléatoire des données
  - **Indépendance** des observations (**pas** mesure répété)
  - **Normalité** des données (pour tous les échantillons ou groupes)
  - **Homoscédasticité** que l'on vérifie avec un test de Bartlett
  - Les résidus e_{ij} doivent suivre une distribution normale, être indépendants et de même variance (estimée par S_e^2)

Analyse de variances

- Présentation des données

Hypothèses :

H_0 : les moyennes des **a** échantillons ne sont pas significativement différentes

H_1 : au moins 1 moyenne est différente des autres

On calcule la **moyenne globale**

$$\bar{X} = \frac{\sum_{i-1}^{a}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{a} n_i}$$

	Echantillon 1	…	Echantillon i	…	Echantillon a
	x_{11}	…	x_{i1}	…	x_{a1}
	⋮		⋮		⋮
	x_{1j}	…	x_{ij}	…	x_{aj}
	⋮		⋮		⋮
	x_{1n_1}	…	⋮	…	x_{an_a}
			x_{in_i}		
Moyenne	$\bar{X}_1$		$\bar{X}_i$		$\bar{X}_a$

Analyse de variances

- Présentation des données

On calcule la **dispersion totale** : SCE_T

(somme des carrés des écarts de l'ensemble des données observées à la moyenne globale, sans tenir compte de l'appartenance des données à un échantillon)

$$SCE_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{X}\right)^2$$

	Echantillon 1	…	Echantillon i	…	Echantillon a
	x_{11}	…	x_{i1}	…	x_{a1}
	⋮		⋮		⋮
	x_{1j}	…	x_{ij}	…	x_{aj}
	⋮		⋮		⋮
	x_{1n_1}	…	⋮	…	x_{an_a}
			x_{in_i}		
Moyenne	$\bar{X}_1$		$\bar{X}_i$		$\bar{X}_a$

On peut décomposer en la somme des carrés des écarts factorielle (SCE_A, entre la moyenne de chaque éch et la moyenne globale) et la somme des carrés des écarts résiduels (SCE_R, entre les valeurs individuelles et la moyenne de son groupe)

$$\boldsymbol{SCE_T = SCE_R + SCE_A}$$

$$SCE_R = \sum_{i-1}^{a}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \qquad SCE_A = \sum_{i-1}^{a}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2$$

Analyse de variances

- Présentation des données :

A partir des dispersions, on calcule des **estimateurs de la variance** (carrés moyens) par la somme des carrés des écarts divisés par le degré de liberté.

$$CM = \frac{SCE}{nb\ ddl}$$

- La **variance totale** : $\boldsymbol{CM_T = \frac{SCE_T}{N-1}}$
- La **variance factorielle** (variance inter-groupe) : $\boldsymbol{CM_A = \frac{SCE_A}{a-1}}$
- La **variance résiduelle** (variance intra-groupe) : $\boldsymbol{CM_R = \frac{SCE_R}{N-a}}$
- => 3 estimateurs de la variance

Analyse de variances

- Statistique du test :

Si H_0 est vraie : La variance intra-groupe (fluctuation d'échantillonnage) n'est significativement pas différente de la variance inter-groupe (fluctuation due au facteur)

  - pas d'effet du facteur

- Test de comparaison de variances -> calcul la variable de décision F.

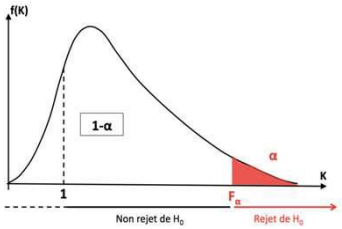
Test obligatoirement **unilatéral** (on suppose que la variance inter-groupe sera plus forte si l'effet du facteur existe => toujours au numérateur)

Analyse de variances

- Statistique du test : $F = \frac{CM_A}{CM_R}$

- Sous H_0 , F (ou K) doit suivre une loi de Fischer-Snedecor à a-1 et a n-a degrés de liberté ($F_{(a-1;\ n-a)}$)

- **Règle de décision :**
- Si $F \geq F_\alpha$ alors H_0 est rejeté,
- alors que si $F < F_\alpha$ alors H_0 est non rejeté



Analyse de variances

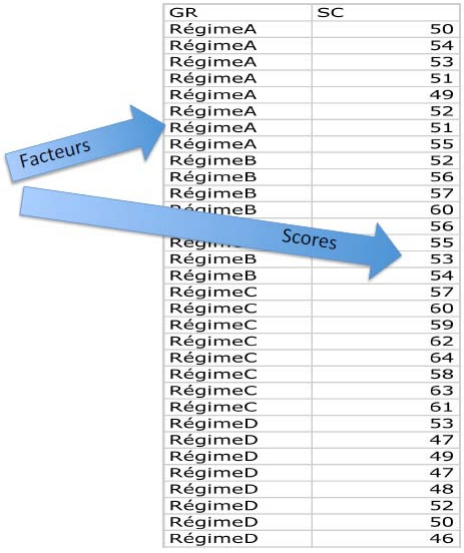
- Tableau de base pour l'ANOVA à 1 facteur :

RégimeA	RégimeB	RégimeC	RégimeD
50	52	57	53
54	56	60	47
53	57	59	49
51	60	62	47
49	56	64	48
52	55	58	52
51	53	63	50
55	54	61	46

- Difficilement exploitable sous R (comme sur pratiquement tous les logiciels de stats)

Analyse de variances

- Le bon format de représentation des données
- Mais on va le simplifier :
  - Colonnes : Régimes et Scores
  - Et Facteurs : RA, RB, RC, RD



GR	SC
RégimeA	50
RégimeA	54
RégimeA	53
RégimeA	51
RégimeA	49
RégimeA	52
RégimeA	51
RégimeA	55
RégimeB	52
RégimeB	56
RégimeB	57
RégimeB	60
	56
	55
RégimeB	53
RégimeB	54
RégimeC	57
RégimeC	60
RégimeC	59
RégimeC	62
RégimeC	64
RégimeC	58
RégimeC	63
RégimeC	61
RégimeD	53
RégimeD	47
RégimeD	49
RégimeD	47
RégimeD	48
RégimeD	52
RégimeD	50
RégimeD	46

Analyse de variances

- Script :

```
# ANOVA 1 facteur #
#########################
donnees<-read.table(file.chose(),h=T)
attach(donnees)
par(mfrow=c(3,2))
stripchart(SC~GR,ver9cal=T,method="jitter",col=c("red","blue","green","brown"))
boxplot(SC~GR,col=c("red","blue","green","brown"))
```

Nouvelle commande “fonction” de R :

```
moyenne<-tapply(SC,GR,mean)
barplot(moyenne,col=c("red","blue","green","brown"))
```

Analyse de variances

- **Script :**

resultat<-aov(SC~GR)

summary(resultat)

$$F = \frac{CM_A}{CM_R} = \frac{196.04}{5.67} = 34.577$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
GR	3	588.13	196.04	34.577	1.491e-09 ***
Residuals	28	158.75	5.67		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- **Tests supplémentaires :**

group <- unique(donnees$GR)

shapiro.test(donnees[donnees$GR==group[i], 'SC']) # Normalité pour tt les groupes

bartlett.test(donnees) # Homoscédasticité

- **Tests alternatives :**

kruskal.test(SC~GR) # non-paramétrique

- mesure répété (ex. même individu plusieurs 'TimePoints' ou 'Treatments'))

aov(SC~TimePoint + Error(GR/TimePoint)) # paramétrique

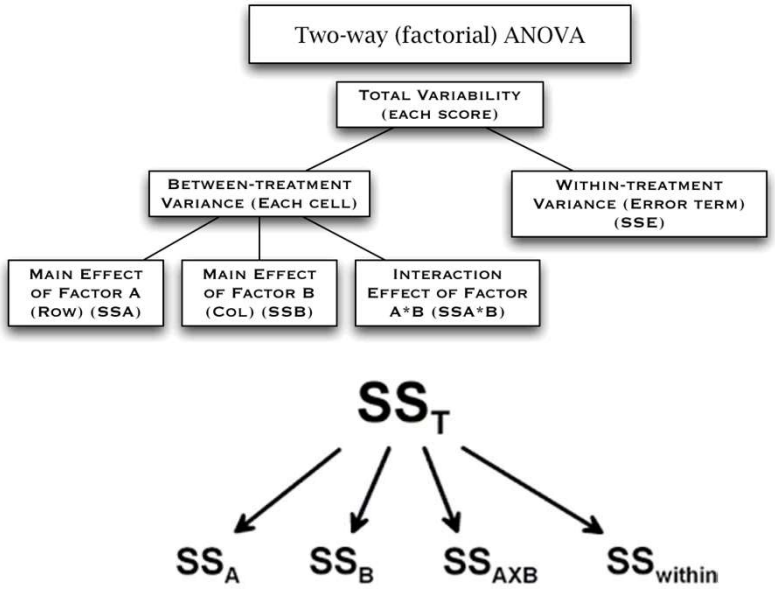
friedman.test(SC~TimePoint|GR) # non-paramétrique

Two-way ANOVA

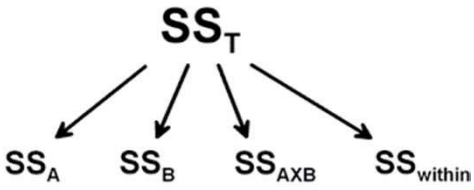
LICENCE SCIENCES ET TECHNOLOGIE

MENTION SCIENCE DE LA VIE – L2

Analyse de variances



Analyse de variances



$$SSR = nC\sum_{i=1}^{R}(\overline{X}_i - \overline{X})^2 \qquad df_R = R-1$$

$$SSC = nR\sum_{j=1}^{C}(\overline{X}_j - \overline{X})^2 \qquad df_C = C-1$$

$$SSI = n\sum_{i=1}^{R}\sum_{j=1}^{C}(\overline{X}_{ij} - \overline{X}_i - \overline{X}_j + \overline{X})^2 \qquad df_I = (R-1)(C-1)$$

$$SSE = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}(X_{ijk} - \overline{X}_{ij})^2 \qquad df_E = RC(n-1)$$

$$SST = \sum_{c=1}^{C}\sum_{r=1}^{R}\sum_{a=1}^{n}(X_{ijk} - \overline{X})^2 \qquad df_T = N-1$$

$$MSR = \frac{SSR}{R-1} \qquad F_R = \frac{MSR}{MSE}$$

$$MSC = \frac{SSC}{C-1} \qquad F_C = \frac{MSC}{MSE}$$

$$MSI = \frac{SSI}{(R-1)(C-1)} \qquad F_I = \frac{MSI}{MSE}$$

$$MSE = \frac{SSE}{RC(n-1)}$$

where:

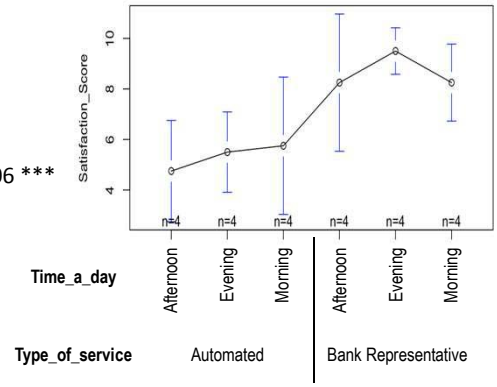
- n = number of observations per cell
- C = number of column treatments
- R = number of row treatments
- i = row treatment level
- j = column treatment level
- k = cell member
- X_{ijk} = individual observation
- $\overline{X}_{ij}$ = cell mean
- $\overline{X}_i$ = row mean
- $\overline{X}_j$ = column mean
- $\overline{X}$ = grand mean

Time_a_day	Type_of_service	Satisfaction_Score
Morning	Automated	6
Morning	Automated	5
Morning	Automated	8
Morning	Automated	4
Morning	Bank Rep	8
Morning	Bank Rep	7
Morning	Bank Rep	9
Morning	Bank Rep	9
Afternoon	Automated	3
Afternoon	Automated	5
Afternoon	Automated	6
Afternoon	Automated	5
Afternoon	Bank Rep	9
Afternoon	Bank Rep	10
Afternoon	Bank Rep	6
Afternoon	Bank Rep	8
Evening	Automated	5
Evening	Automated	5
Evening	Automated	7
Evening	Automated	5
Evening	Bank Rep	9
Evening	Bank Rep	10
Evening	Bank Rep	10
Evening	Bank Rep	9

Analyse de variances

```
> resultat<-aov(Satisfaction_Score~Time_a_day*Type_of_service)
> summary.aov(resultat)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Time_a_day	2	4.00	2.00	1.241	0.313	
Type_of_service	1	66.67	66.67	41.379	4.7e-06	***
Time_a_day:Type_of_service	2	2.33	1.17	0.724	0.498	
Residuals	18	29.00	1.61			




```
> plotmeans(Satisfaction_Score ~ interaction(Time_a_day, Type_of_service, sep = " "))
```

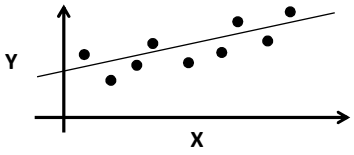
ANCOVA

LICENCE SCIENCES ET TECHNOLOGIE

MENTION SCIENCE DE LA VIE – L2

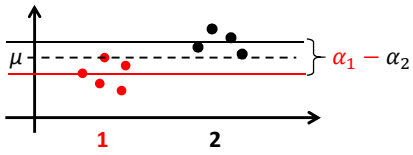
Regression lineare + analyse de variances

Linear regression



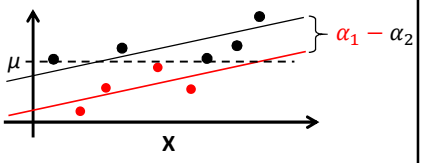
$$y = \mu + \beta * x + \varepsilon$$

ANOVA



$$y = \mu + \begin{pmatrix} Treat & Coeff \\ 1 & \alpha_1 \\ 2 & \alpha_2 \end{pmatrix} + \varepsilon$$

ANCOVA



$$y = \mu + \begin{pmatrix} Treat & Coeff \\ 1 & \alpha_1 \\ 2 & \alpha_2 \end{pmatrix} + \beta * x + \varepsilon$$

Corrélation et régression linéaire

LICENCE SCIENCES ET TECHNOLOGIE
MENTION SCIENCE DE LA VIE – L2

Corrélation linéaire

- Définitions :
  - Liaison entre plusieurs variables : étude de la distribution conjointe de 2 VA simultanées X et Y
- Quelles VA ?

  2 VA **quanti** distribuées **normalement**

  => Corrélation linéaire simple

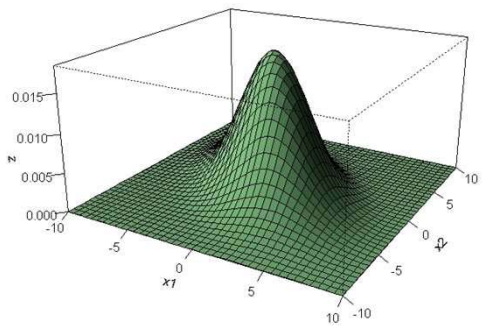
  2 VA **simultanées** (sur les mêmes individus d'un échantillon)

  2 VA dont les mesures sont **indépendantes**

  2 VA **corrélées** (ou non)

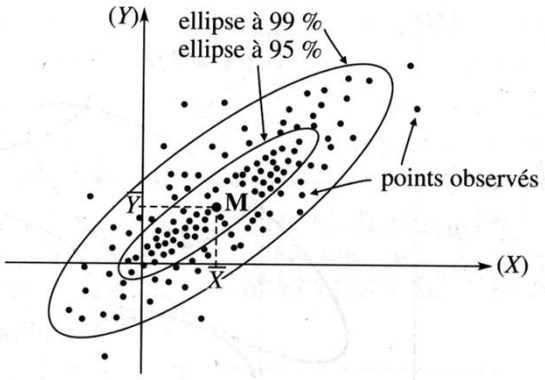
Corrélation linéaire

- On construit le vecteur aléatoire [X;Y] de dimension 2 = **nouvelle VA binormale**



Corrélation linéaire

- Exemple de distribution d'une bivariable



Corrélation linéaire

- La distribution binormale
  - Distribution normale pour chaque variable (distribution marginales de X et de Y)

  VA normale X -> $N(\mu_x, \alpha_x^2)$

  VA normale Y -> $N(\mu_y, \alpha_y^2)$

  => variances marginales
  - Liaison entre les deux variables : ρ = coefficient de corrélation linéaire
  - Point moyen théorique de la distribution de la bivariable (X, Y) : μ (coordonnées : (μ_x, μ_y))

Corrélation linéaire

- La covariance :
  - Extension du concept de variance à deux VA

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2 \quad \left| \quad \hat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu}_y)^2 \right.$$

  - Mesure la dispersion d'un nuage de points dans un espace à deux dimensions (X, Y)

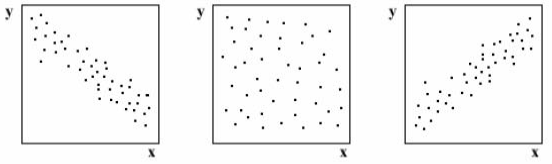
$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)\right]$$

$$Cov(x,y) = \frac{n\sum xy - \sum x \sum y}{n(n-1)}$$

Corrélation linéaire

- La covariance
  - Mesure la **dispersion *conjointe de deux variables. La covariance*** renseigne sur la forme et l'orientation du nuage de points d'un diagramme de dispersion.

Contrairement à la variance qui ne peut-être que positive, la covariance peut-être **négative, nulle ou positive.**



Cov(x,y) < 0 Cov(x,y) = 0 Cov(x,y) > 0

Corrélation linéaire

- **Attention** : la covariance renseigne sur **l'inclinaison du nuage de points** mais elle ne donne aucune idée de l'intensité de la liaison existant entre les 2 variables.
  - ⇒**Trouver une autre mesure qui renseigne sur l'intensité de la relation.**
  - ⇒ **Mesure de la corrélation paramétrique = r de Pearson**

- **A garder en tête :**
  - **Covariance : mesure de dispersion conjointe de 2 variables quantitatives autour** de leur moyenne.
  - **Corrélation : mesure de la liaison entre 2 variables.**

Corrélation linéaire

- Coefficient de corrélation :

La covariance dépend des unités et des ordres de grandeurs des VA mesurées.

=> pour s'en affranchir : covariance des **VA centrées réduites**

=> Coefficient de corrélation **ρ**

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Sur l'échantillon :

$$Cov(x', y') = \frac{Cov(x,y)}{\hat{\sigma}_x \hat{\sigma}_y} = r_{xy}$$

Corrélation linéaire

- Coefficient de corrélation de Pearson (r) :

$$r_{xy} = \frac{Cov(x,y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\sqrt{\sum_{i=1}^{n} (x_i - \hat{\mu}_x)^2 \sum_{i=1}^{n} (y_i - \hat{\mu}_y)^2}}$$

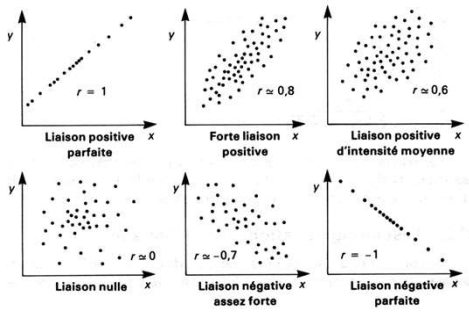
- Propriétés : r E [-1;1]
  - 1.

	x	y
x	$r_{xx}=1$	r_{xy}
y	r_{xy}	$r_{xy}=1$

Les valeurs des diagonales sont « 1 » puisque les variables sont centrées et réduites.

Corrélation linéaire

- Propriétés du coefficient de corrélation de Pearson (r) :
  - 2. r= +1 ou r= -1 si les points forment une ligne droite dans le diagramme de dispersion
  - 3. Le signe de r est le même que le signe de la covariance, il indique si la relation est de pente positive (croissante) ou négative (décroissante).



Corrélation linéaire

- Remarque sur le coefficient de corrélation r :

La covariance d'un échantillon = estimateur sans biais de la covariance de la population

=> Le coefficient de corrélation d'un échantillon = estimateur sans biais du coefficient de corrélation de la population

Corrélation linéaire

- **Test statistique sur r :**

Pour savoir s'il y a corrélation ou pas entre les 2 variables, il faut pouvoir juger de la différence de r_{xy} par rapport à 0.

En d'autres termes, r_{xy} est-il significativement différent de zéro ou pas ?

S'il est significativement différent de 0, il y a corrélation entre les variables.

S'il n'est pas significativement différent de 0, il n'y a pas corrélation entre les variables.

Question : le coefficient de corrélation calculé r_{xy} est-il significativement différent de 0 ? Les variables x et y sont-elles liées ?

Corrélation linéaire

- **Test statistique sur r :**

Hypothèses :

H_0 : ρ = 0

H_1 : ρ ≠ 0 test bilatéral **ou** H_1 : ρ > 0 si liaison positive ou ρ < 0 si liaison négative

Conditions d'application :

- Les 2 variables sont quantitatives
- La distribution dans la population est bi-normale.
- Les observations sont indépendantes.

Corrélation linéaire

- **Test statistique sur r :**

Distribution de la statistique sous H0 : T suit une distribution de Student à **n-2** ddl.

$$t = \frac{r_{xy}}{\sqrt{(1 - r_{xy}^2)/(n-2)}}$$

- **Règle de décision :**

Soit $|T| > t_{\alpha/2}$, alors H_0 rejetée au risque α

Soit $|T| < t_{\alpha/2}$, alors H_0 non rejetée au risque α

En unilatéral : $|T| > t_\alpha$ -> H_0 rejetée au risque α

$|T| < t_\alpha$ -> alors H_0 non rejetée au risque α

- Attention, le test répond à la question de liaison significative ou non entre les 2 variables.

Corrélation linéaire

- **Test statistique sur r :**

Pour un test bilatéral : utilisation de la table de signification du r de Bravais-Pearson

$|r_c| < r_\alpha$: Non-rejet de H_0

$|r_c| \geq r_\alpha$: Rejet de H_0

- Attention : si corrélation : pas de relation de cause à effet !!

Corrélation linéaire

- Sous R :

```
longueur=c(7.4,7.7,8.2,8,9,9.4,9.5,9.1,9.7,8.5,9.3,9.6,8.4,7.8,8.6)
ovocytes=c(25,41,47,46,58,73,89,79,78,60,85,93,67,37,53)
plot(ovocytes,longueur)
cor.test(longueur,ovocytes)
```

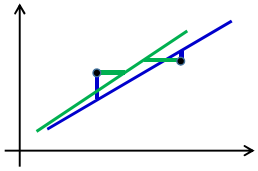
Régression linéaire

- La corrélation montre qu'il existe un lien entre les variables mais ne le caractérise pas.

  ⇒La méthode de la régression a pour but de décrire la relation entre 1 **variable aléatoire quantitative (Y) dite variable dépendante et une variable quantitative (X) dite variable explicative.**

Si la variable X est **contrôlée, on parle de régression de *modèle I (Y~X)*.**

Si la variable X est **aléatoire, on parle de régression de *modèle II (X~Y)*.**



S'il y a **plusieurs variables X explicatives, on parle de régression multiple** ($y = ax_1 + bx_2 + cx_3 + \ldots + d$)

Si le problème de régression n'implique qu'une seule variable explicative et simplement au premier degré (non x_2, x_3…), il s'agit de régression linéaire simple.

Régression linéaire

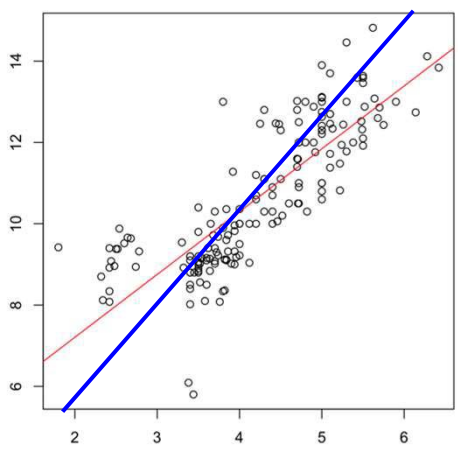
- Ici, on étudie **la régression linéaire simple de modèle I c'est-à-dire celle correspondant à une équation de type y=a x+b**

Cette droite porte le nom de **droite d'estimation ou droite de régression de y en x.**
(On peut considérer également la droite de régression de x en y – modèle II)

On travaille avec :
2 VA **quantitatives** X et Y distribuées **normalement** (distribution binormale)
2 VA **simultanées**
2 VA **dont les mesures sont indépendantes**
2 VA **corrélées** entre elles significativement

Régression linéaire

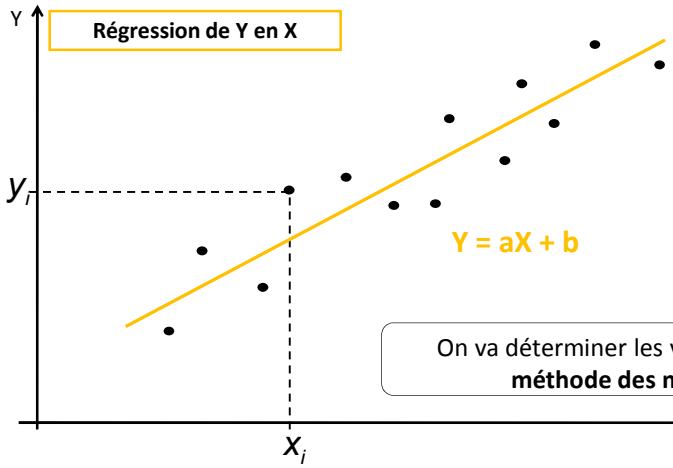
Un nuage de points :
2 droites de régression



Régression de X en Y
X = cY + d

Régression de Y en X
Y = aX + b

Régression linéaire



Y = aX + b

a = pente de la droite de régression

b = ordonnée à l'origine

On va déterminer les valeurs de a et b par la **méthode des moindres carrés**

Régression linéaire

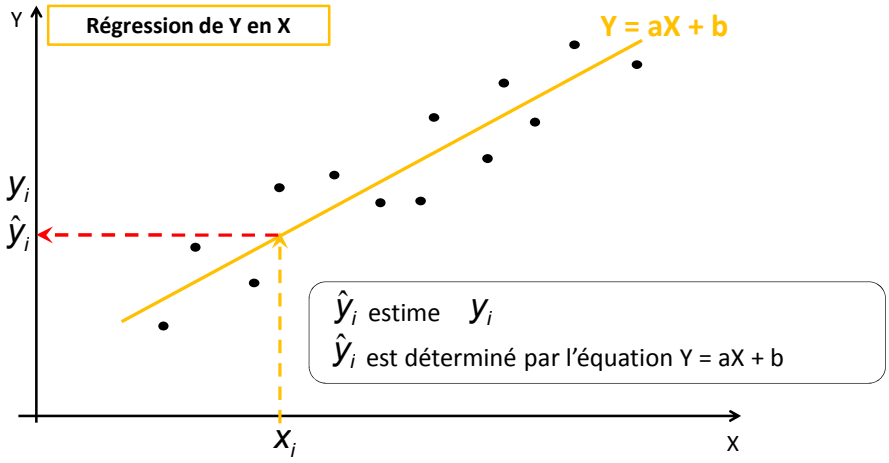
- Dans un premier temps : décrire les données par un modèle linéaire
- On cherche à prédire les valeurs de Y pour chaque valeur de X.
  ⇒Construire un **modèle statistique** simple qui prend la forme suivante :

$$y_i = \hat{y}_i + \varepsilon_i$$

Prédiction de y faite pour l'individu i et avec ε_i une VA correspondant à la différence entre la valeur observée et la valeur prédite de y.

ε_i = VA décrivant l'**erreur** dans la prédiction ou **résidu**

Régression linéaire



Régression linéaire

Si l'on suppose qu'il existe une relation linéaire entre x et y, on peut écrire :

$$\hat{y}_i = \hat{\mu}_y + \alpha\left(x_i - \hat{\mu}_x\right)$$

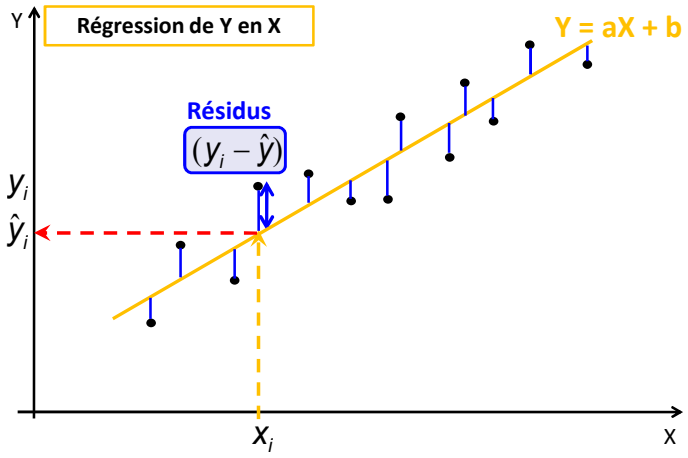
- On peut calculer l'erreur globale :

  => Somme des carré des écarts résiduels : SCER

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$SCER = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (\hat{\mu}_y - y_i + \alpha\left(x_i - \hat{\mu}_x\right))^2$$

Régression linéaire



Régression linéaire

- Déterminer la droite qui passe au mieux dans le nuage de points = **la droite qui globalement est la plus proche de l'ensemble des points observés.**
  ⇒Droite minimisant la SCER.
- On cherche alors la valeur de $\hat{\alpha}$ de α minimisant la SCER, par la méthode des moindres carrés pour trouver l'équation de la droite qui s'ajuste au mieux au nuage de points.
- Cette valeur est telle que $\delta SCER / \delta \alpha = 0$
- Le résultat est :

$$\hat{\alpha} = \frac{\sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\sum (x_i - \hat{\mu}_x)^2} = \frac{\text{cov}(x,y)}{\hat{\sigma}_x^2} = r_{xy}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$$

Régression linéaire

- Régression de Y en X :

  Y = aX + b

$$a = r\frac{S_Y}{S_X}$$

$$b = \bar{y} - a\bar{x}$$

- Régression de X en Y :

  X = cY + d

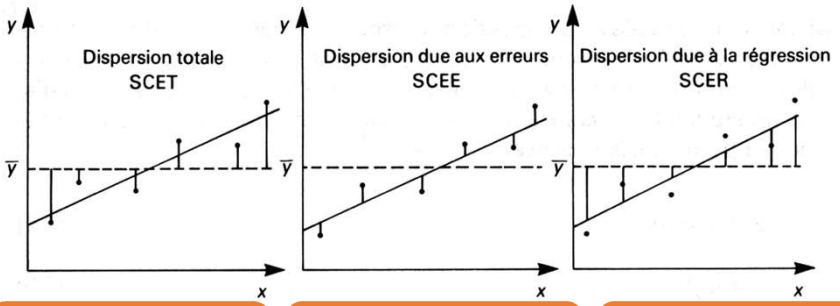
$$c = r\frac{S_X}{S_Y}$$

$$d = \bar{x} - c\bar{y}$$

Régression linéaire

- Coefficient de détermination R^2 et décomposition de la variance
  - Test de Fisher sur les variances (expliquée et résiduelle)
  - Coefficient de détermination R^2 : mesure de la proportion de la variation de Y expliquée par la régression (donc par la variation de X)
  - Décomposition de la variance de Y (transposable sur X)

Régression linéaire



$$SCE_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SCE_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad SCE_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SCE_T = SCE_E + SCE_R$$

Régression linéaire

- Coefficient de détermination R² : $R^2 = \frac{SCE_R}{SCE_T}$

= proportion de la variation de Y expliquée par la régression.

Plus R² est proche de 1, meilleure est la prédiction fournie par le modèle.

Dans la régression linéaire simple : $R^2 = \hat{r}_{xy}^2$

Régression linéaire

- Analyse de la liaison entre 2 VA X et Y :
  - Coefficient de corrélation r = indication de la signification de la corrélation (en fonction de la taille de l'échantillon)
  - Coefficient de détermination R² = proportion de la variance expliquée par la régression
- **Remarques importantes :** $\sqrt{R^2} = r$
  - **Une valeur significative de r n'implique pas qu'une grande partie de la variance totale de Y est expliquée par la régression**
  - **Un fort R² n'implique pas une corrélation significative**

Régression linéaire

- **Test de Fisher sur les variances (expliquée et résiduelle)**

Pour tester la significativité du modèle, on peut s'intéresser directement au rapport F, entre SCEE et SCER

$$F = (n-2)\frac{SCE_R}{SCE_E}$$

- Condition de validité :

Les résidus doivent être normalement distribués.

- Hypothèses :

H_0 : le modèle n'est pas explicatif (variance expliquée = 0, $\rho^2 = 0$)

H_1 : le modèle est explicatif (variance expliquée>0, $\rho^2 \neq 0$)

Régression linéaire

- **Test de Fisher sur les variances (expliquée et résiduelle)**

- Statistique : $f_{obs} = (n-2)\frac{SCE_R}{SCE_E}$

F suit une loi de Fisher à 1 (SCE_R) et n-2 (SCE_E) ddl

- Règle de décision :

Si $F \geq F_\alpha$ alors H_0 est rejeté,

alors que si $f < F_\alpha$ alors H_0 est non rejeté

Régression linéaire

- En régression linéaire simple le **test t de Student sur la pente de la régression** $\hat{\alpha}$ est équivalent au test de Fisher avec la relation suivante : Fc = t^2.

- Les valeurs de p des deux tests sont égales.

Si les résidus sont normalement distribués, $\hat{\alpha}$ suit une distribution normale centre dont la variance est estimée par :

$$\hat{\sigma}^2_{\hat{\alpha}} = \frac{\sum(y_i - \hat{y}_i)^2}{(n-2)\sum(x_i - \bar{x})^2} = \frac{SCE_E}{(n-2) * SCEX}$$

- On utilise ce résultat pour tester la significativité de la relation entre X et Y.

Régression linéaire

- Condition de validité :

Les résidus doivent être distribués normalement.

- Hypothèses :

H_0 : la pente de la droite de régression est nulle : $\alpha = 0$

H_1 : $\alpha \neq 0$ (la pente n'est pas nulle) en bilatéral ou $\alpha < 0$ et $\alpha > 0$ en unilatéral.

- Statistique :

T suit une distribution de Student à n-2 ddl $t_{obs} = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$

- **Règle de décision :**

Soit $|T| > t_{\alpha/2}$, alors H_0 rejetée au risque α

Soit $|T| < t_{\alpha/2}$, alors H_0 non rejetée au risque α

En unilatéral : $|T| > t_\alpha$ -> H_0 rejetée au risque α

$|T| < t_\alpha$ -> alors H_0 non rejetée au risque α

Régression linéaire

- Sous R :

```
> longueur=c(7.4,7.7,8.2,8,9,9.4,9.5,9.1,9.7,8.5,9.3,9.6,8.4,7.8,8.6)
> ovocytes=c(25,41,47,46,58,73,89,79,78,60,85,93,67,37,53)
# Normalité
> shapiro.test(ovocytes)
> shapiro.test(longueur)
> plot(ovocytes,longueur)
# Correlation
> cor.test(longueur,ovocytes)
# Régression linéaire
> reg1=lm(longueur~ovocytes)
> summary(reg1)
> shapiro.test(reg1$residuals)
> plot(ovocytes, longueur)
> abline(reg1)
# Régression linéaire inverse
> reg2=lm(ovocytes~longueur)
> summary(reg2)
> plot(longueur,ovocytes)
> abline(reg2)
```