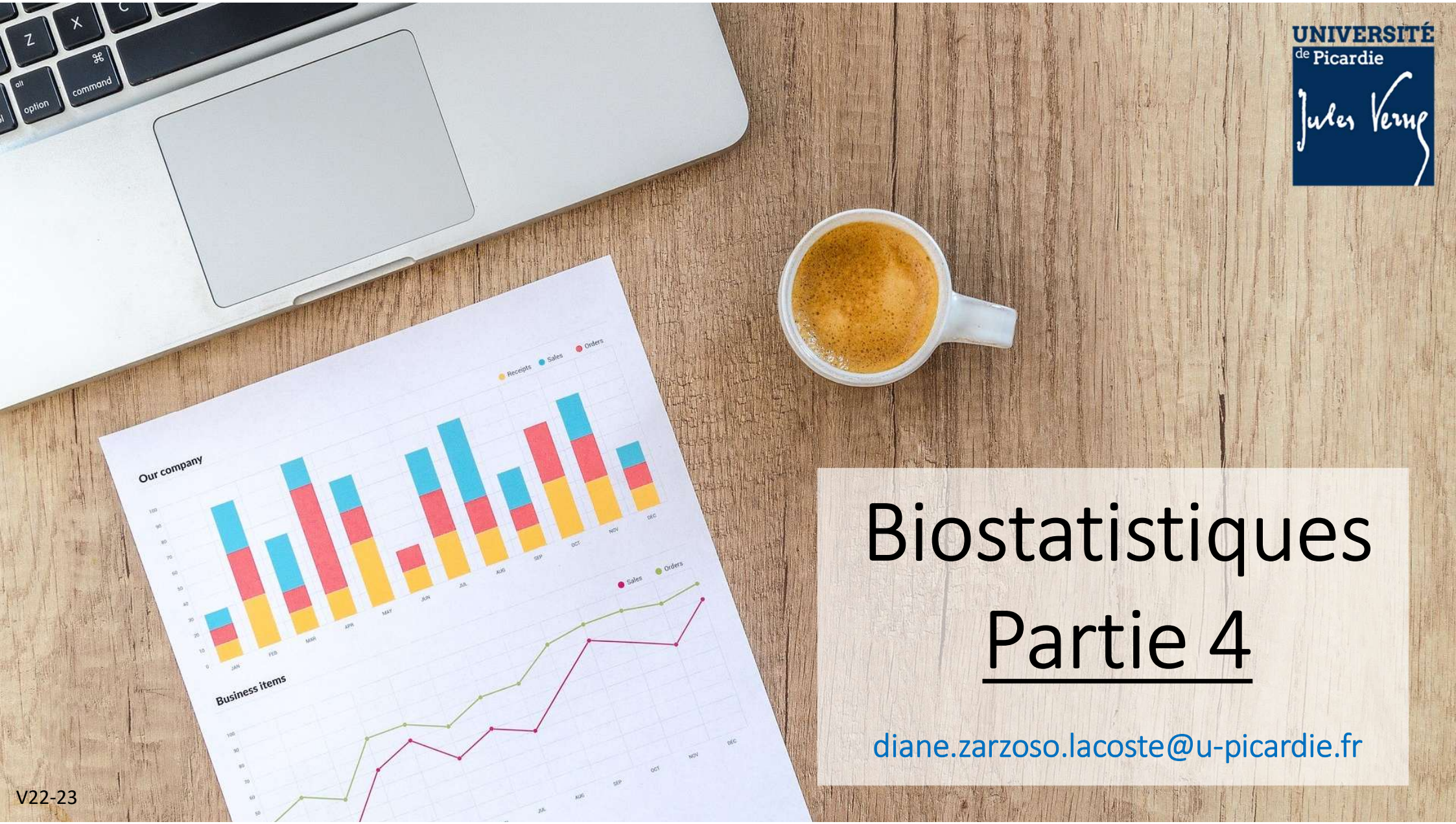
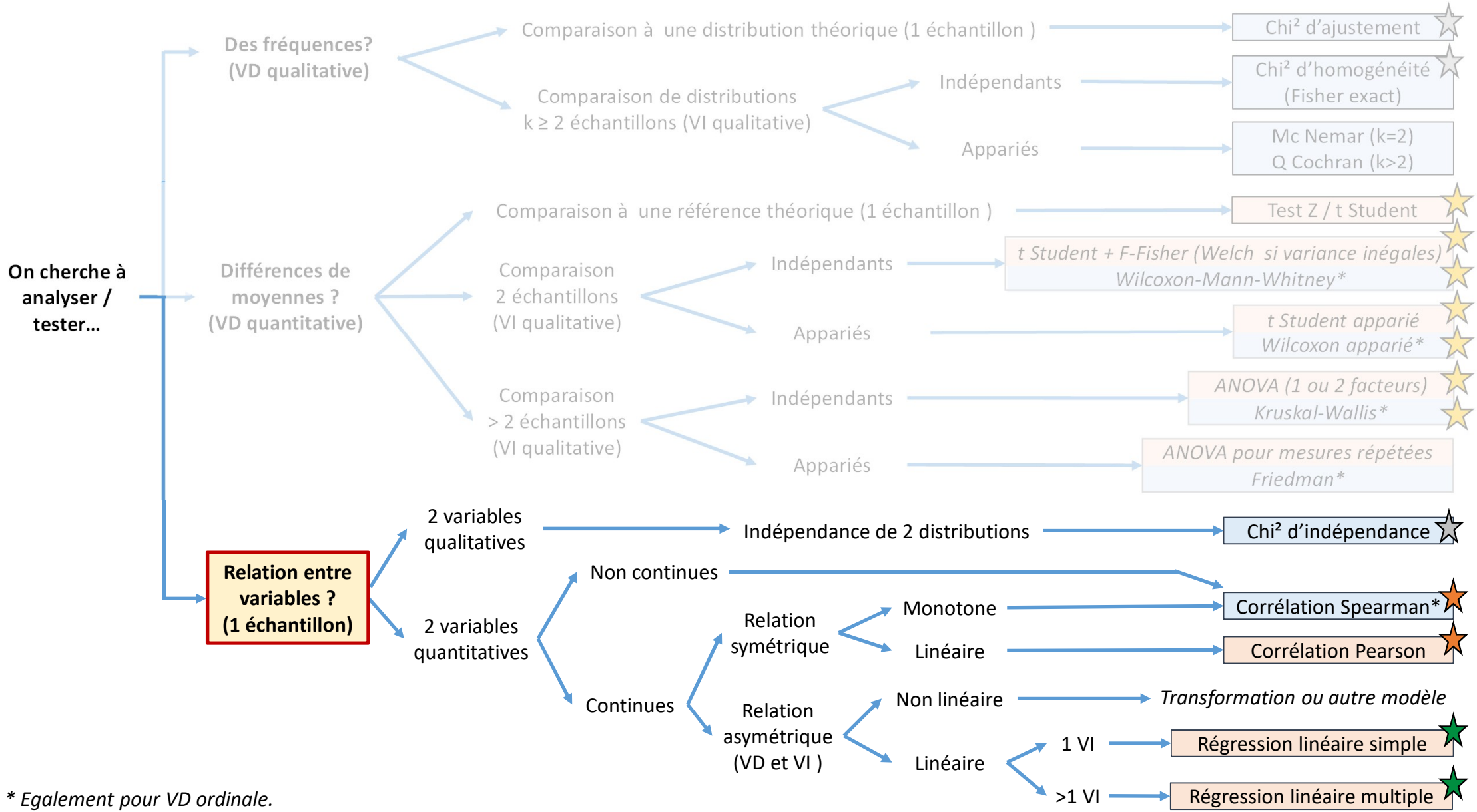


UNIVERSITÉ de Picardie Jules Verne

Biostatistiques
Partie 4

diane.zarzoso.lacoste@u-picardie.fr

V22-23



* *Egalement pour VD ordinale.*

Etude de la relation entre 2 variables - Introduction

Analyse bivariée : Etude **simultanée** de 2 VA, X et Y, définies sur 1 même population d'individus.

→ 1 échantillon de taille N, avec, pour chaque individu i, 1 couple d'observations (X_i, Y_i).

Individu i	X	Y
1	X_1	Y_1
2	X_2	Y_2
3	X_3	Y_3
…	…	…
N	X_N	Y_N

But : Rechercher et quantifier une éventuelle liaison/dépendance entre les 2 variables .

Les 2 variables sont elles liées/dépendantes ? les variations de l'une dépendent des variations de l'autre → *Sachant que l'on connaît la valeur prise par une variable pour un individu, peut-on avoir une indication sur la valeur prise par l'autre variable sur cet individu ?*

Ou indépendantes ? la connaissance de la valeur prise par l'1 des 2 variables sur 1 individu n'apporte aucune information sur la valeur prise par l'autre variable.

Etude de la relation entre 2 variables - Introduction

Importance du rôle des variables dans l'étude de leur relation :

Rôle symétrique: les variables sont **interchangeables**, on étudie leur **co-variation sans hypothèse causale** (elles varient conjointement, mais l'une n'explique pas l'autre).

- 2 VA qualitatives (ou quantitatives en classes) → **test d'indépendance (Chi²)**

  Ex: la couleur des yeux et la couleur des cheveux.

- 2 VA quantitatives → **test de corrélation (Pearson, Spearman)**

  Ex: Taille et masse des individus, longueur des pattes et vitesse de déplacement

Etude de la relation entre 2 variables - Introduction

Importance du rôle des variables dans l'étude de leur relation :

Rôle asymétrique: On cherche à expliquer/prédire les variations d'1 VD, par celle d'1 VI.
→ relation causale = 1 variable agit sur 1 autre (lien directionnel).

- 1 VD quantitative et 1 VI discrète/qualitative → **Anova à un facteur**

  Ex: Influence du type d'engrais (VI) sur le rendement en blé (VD)

- 1 VD et 1 VI quantitatives → **régression (linéaire) simple**

  Ex: Expliquer et prédire la masse des individus par leur taille
  Expliquer et prédire la vitesse de déplacement par la longueur des pattes

- 1 VD et >1 VI quantitatives → **régression (linéaire) multiple**

Ex: Expliquer et prédire la vitesse de déplacement par la longueur des pattes et masse

Plan du cours

Partie 5. Etude des relations entre variables

5.1. Le coefficient de corrélation

5.1.1. Coefficient de corrélation de Pearson

5.1.2. Coefficient de corrélation de Spearman

5.2. De la corrélation au modèle de régression

5.2.1. Régression linéaire simple

5.2.2. Régression linéaire multiple

Le coefficient de corrélation pour 2 variables quantitatives

Soient ***X* et *Y*** 2 VA **quantitatives symétriques** observées simultanément sur N individus.

Objectifs :

1. Déterminer s'il existe une relation statistiquement significative entre *X* et *Y*.

→ *X* et *Y* sont corrélées si elles tendent à varier l'une en fonction de l'autre.

2. Caractériser la **forme** (linéaire, monotone ..), le **sens** (positive ou négative), et **l'intensité** (forte, faible) de la relation entre *X* et *Y* .

- **Corrélation positive** : les 2 VA tendent à varier dans le même sens.
- **Corrélation négative** : les 2 VA tendent à varier en sens contraire.

Comment faire?

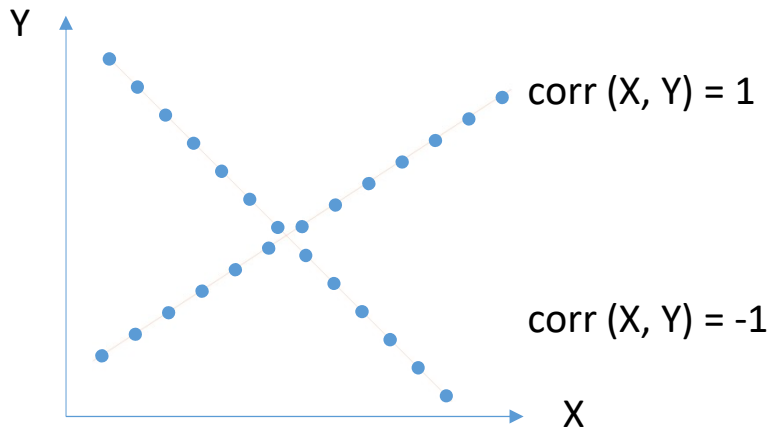
Outil graphique : le nuage de points = Diagramme de dispersion → plot(), pairs()

Indicateur numérique (de sens et intensité) : coefficient de corrélation (r)

Le coefficient de corrélation pour 2 variables quantitatives

Ce coefficient (ρ = pop., r = éch.), **varie entre -1 et 1** et permet de connaitre **l'intensité** (valeur de |r|) et le **sens** (signe de r) de la relation entre X et Y.

Corrélation parfaite: Liaison à 100% entre X et Y



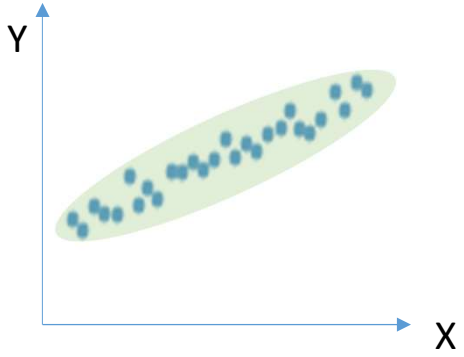
- La connaissance d'une variable donne exactement la valeur de l'autre → elles sont mutuellement déterminées au moyen d'une relation linéaire (forme) du type y = ax + b.
- Cette corrélation parfaite peut être positive (croissante) ou négative (décroissante).

Le coefficient de corrélation pour 2 variables quantitatives

Ce coefficient (ρ = pop., r = éch.), **varie entre -1 et 1** et permet de connaitre **l'intensité** (valeur de |r|) et le **sens** (signe de r) de la relation entre X et Y.

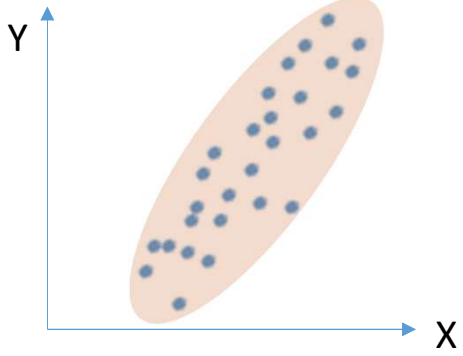
L'intensité de la corrélation est d'autant plus forte que |r| est proche de 1

corr (X, Y) = 0.90



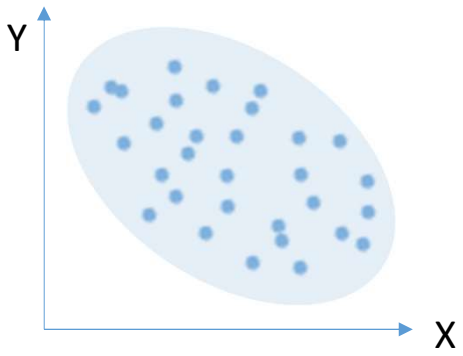
Corrélation linéaire très forte

corr (X, Y) = 0.70



Corrélation forte

corr (X, Y) = - 0.50



Corrélation moyenne

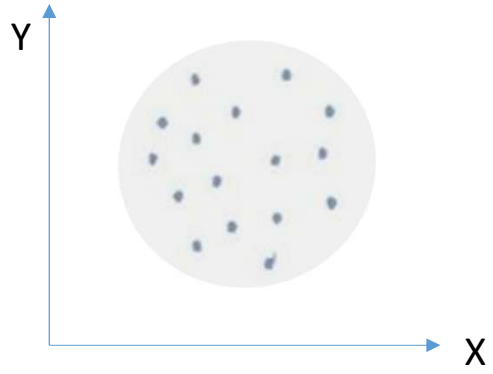
→ En pratique, la corrélation est importante si I r I > 0.6 ou 0.7 (subjectif)

NB: on ne s'intéresse pas à la pente de la relation linéaire entre les 2 variables mais à la dispersion du nuage de point autour de cette relation!!

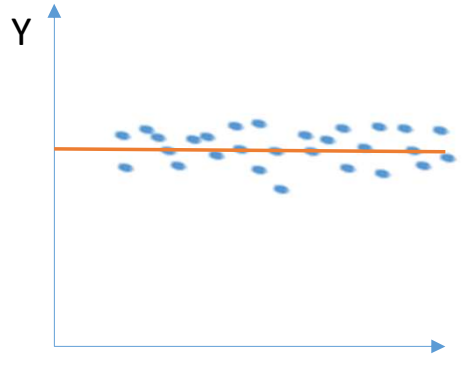
Le coefficient de corrélation pour 2 variables quantitatives

Ce coefficient (ρ = pop., r = éch.), **varie entre -1 et 1** et permet de connaitre **l'intensité** (valeur de |r|) et le **sens** (signe de r) de la relation entre X et Y.

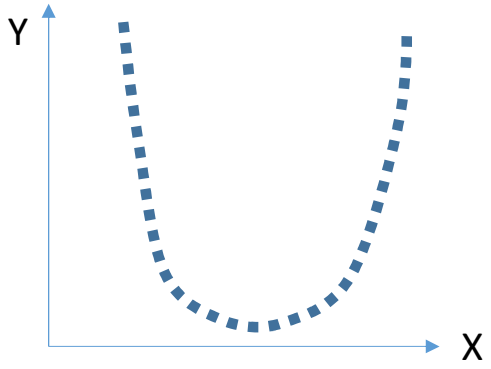
Cas des corrélations nulles: corr (X, Y) ≈ 0



On ne peut pas tracer de droite pour résumer le nuage de point.
Pas de liaison entre X et Y.
→ **indépendance**



On peut tracer une droite mais Y ne varie pas ou peu en fonction de X.
Pas de liaison entre X et Y.
→ **indépendance**



Il existe une liaison entre X et Y, mais elle n'est ni monotone, ni linéaire.

Le coefficient de corrélation pour 2 variables quantitatives

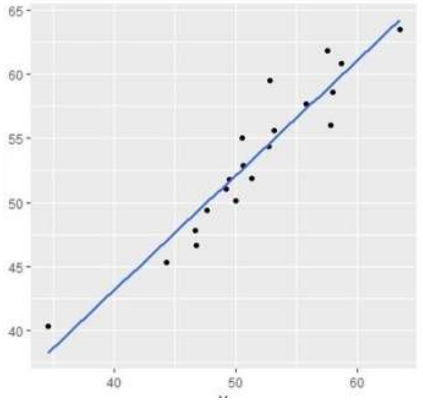
Il existe deux types principaux de coefficients de corrélation: (même interprétation)

5.1.1. Coefficient de Pearson : r

Mesure un lien **linéaire** entre 2 VA quantitatives

Test paramétrique : nécessite distribution **Normale** de **X et Y**

(unimodale symétrique).

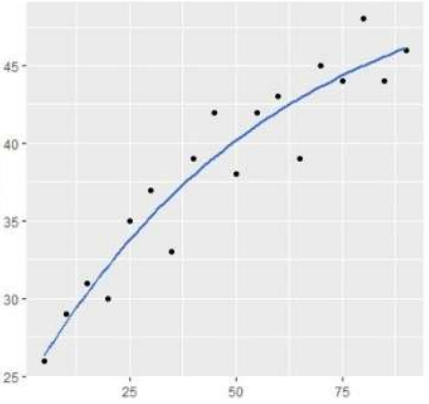


5.1.2. Coefficient de Spearman : r_S

Quantifie la **monotonicité** du lien (**non linéaire**) entre 2 VA exprimées en rangs (VA quantitatives ou ordinales)

Alternative non-paramétrique basée sur les rangs

Corrélation parfaite de ±1 quand une des variables est une fonction monotone parfaite de l'autre (adapté fonctions curvilignes –exponentielle, puissance-)



Le coefficient de corrélation pour 2 variables quantitatives

Test de significativité : est ce que $r_{(s)_XY}$ est significativement ≠ 0 ou non ? (test nullité coeff.)

Condition d'application :

Les observations sont indépendantes (les mesures/individus statistiques ne sont pas liées)

Hypothèses :

- H_0 : r = 0 , il n'existe pas de relation entre X et Y (= indépendance ?)
- H_1 : r ≠ 0, il existe une relation entre X et Y (= liaison/dépendance)

Interprétation :

1. **d'abord la significativité du test de $r_{XY} \neq 0$:** *(manuellement ou sous R)*

   p_{value} → 0.05 > r* = signif. > 0.01 > r** = très signif. > 0.001 > r*** = hautement signif.

2. **ensuite la valeur de r_{XY}** → la corrélation est :
   - positive ou négative (signe de r)
   - forte ou faible (valeur de |r|)

Partie 5. Etude des relations entre variables

5.1. Le coefficient de corrélation

5.1.1. Coefficient de corrélation de Pearson

5.1.2. Coefficient de corrélation de Spearman

5.2. De la corrélation au modèle de régression

5.2.1. Régression linéaire simple

5.2.2. Régression linéaire multiple

Coefficient de corrélation de Pearson

Pour chez vous

Calcul (pour info) :

En pratique, on estime le coefficient de corrélation ρ (pop) grâce à r (ech).

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

Covariance de X et Y: mesure de la dispersion conjointe de 2 variables autour de leur moyenne

Standardisée par le produit des écarts types de X et Y

Pas d'unité

NB: La covariance entre X et Y: - donne le signe du coefficient de corrélation (sens du lien)
- est d'autant plus grande que les variances de X et Y sont grandes

$cov(X,Y)$

Soit :

$$r(X,Y) = \frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{\sum(Xi - \bar{X})^2 * \sum(Yi - \bar{Y})^2}} = \frac{\sum(XiYi) - N\bar{X}\bar{Y}}{\sqrt{(\sum X_i{}^2 - N\bar{X}^2) * (\sum Y_i{}^2 - N\bar{Y}^2)}}$$

Coefficient de corrélation de Pearson

Application:



On a mesuré la richesse en espèces ligneuses et herbacées au sein de 9 forêts tempérées européennes.

Les résultats sont consignés dans le tableau suivant :

RS sp ligneuses	1	2	2	3	4	5	5	6	7
RS espèces herbacées	20	18	25	27	24	28	29	35	40

Existe t il une corrélation entre le nombre d'espèces ligneuses et non-ligneuses au sein de ces forêts?

Effectuer ce test sous R

Copiez le fichier « Diversite_foret.csv » (Moodle) dans votre répertoire de travail, et importer le fichier csv sous R :

Code :

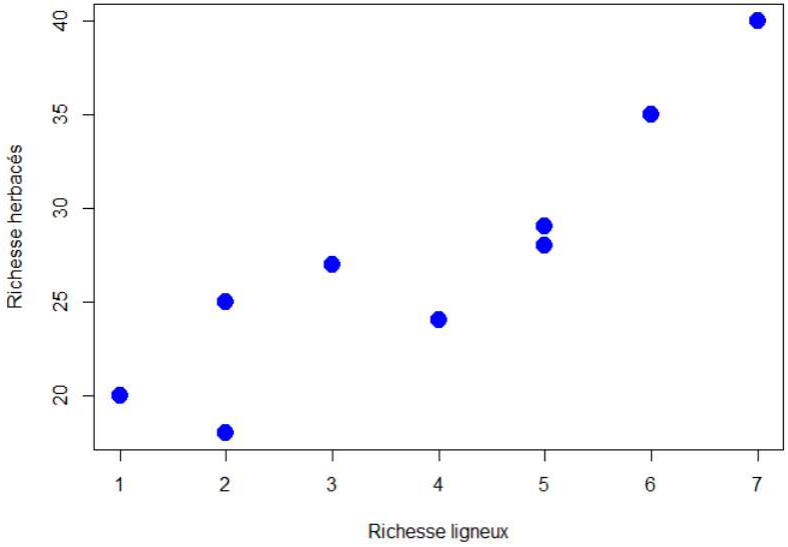
```
Forets<-read.csv("Diversite_foret.csv", sep = ";",h=T)
View(Forets)
str(Forets)

'data.frame':   9 obs. of  2 variables:
 $ RS_Ligneux : int  1 2 2 3 4 5 5 6 7
 $ RS_herbaces: int  20 18 25 27 24 28 29 35 40


plot(Forets$RS_herbaces~Forets$RS_Ligneux,
     pch=20, col="blue", cex=3,
     xlab="Richesse ligneux",
     ylab = "Richesse herbacés")
```

Sorties :

	RS_Ligneux	RS_herbaces
1	1	20
2	2	18
3	2	25
4	3	27
5	4	24
6	5	28
7	5	29
8	6	35
9	7	40



Effectuer ce test sous R

Evaluez graphiquement et à l'aide d'un test la normalité de la distribution des variables.

Code :

```
shapiro.test(Forets$RS_Ligneux)
```

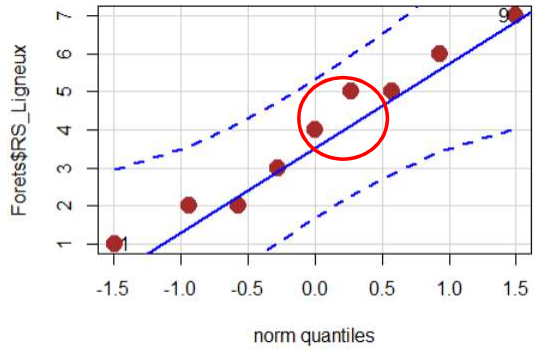
```
library(car)
qqPlot(Forets$RS_Ligneux,
       pch=19,col="brown", cex=2)
```

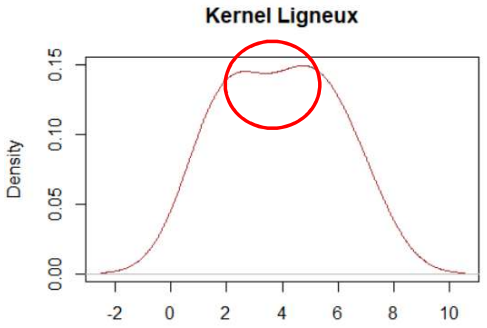
```
kernel_Lignx <- density(Forets$RS_Ligneux)
plot(kernel_Lignx,main="Kernel Ligneux",
     col="brown")
```

Sorties :

```
        Shapiro-Wilk normality test

data:  Forets$RS_Ligneux
W = 0.95823, p-value = 0.7796
```





Effectuer ce test sous R

Code :

```
shapiro.test(Forets$RS_herbaces)
```

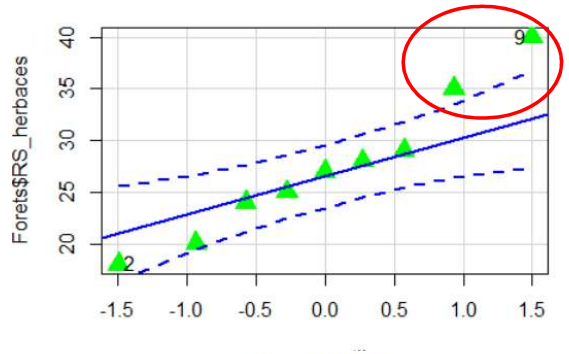
```
qqPlot(Forets$RS_herbaces,
       pch=17,col="green", cex=2)
```

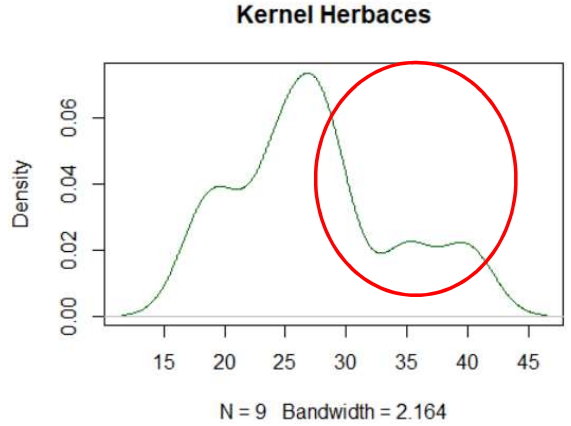
```
hist(Forets$RS_herbaces)
kernel_Herb <- density(Forets$RS_herbaces)
plot(kernel_Herb,main="Kernel Herbaces",
     col="dark green")
```

Sorties :

```
        Shapiro-Wilk normality test

data:  Forets$RS_herbaces
W = 0.95913, p-value = 0.7891
```





Effectuer ce test sous R

Code :

```
cor(Forets$RS_herbaces, Forets$RS_Ligneux)
```

Sorties :

```
[1] 0.9064335
```

NB : Calcul du coefficient de corrélation de Pearson par défaut, ordre des variables importe peu.

```
cor.test(Forets$RS_Ligneux,Forets$RS_herbaces, method="pearson")
```

```
        Pearson's product-moment correlation

data:  Forets$RS_Ligneux and Forets$RS_herbaces
t = 5.6782, df = 7, p-value = 0.0007521
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6087962 0.9803824
sample estimates:
      cor
0.9064335
```

Forte relation linéaire positive entre les 2 variables (r = 0.90), hautement significative.

(Mais non interprétable, car conditions d'application non remplies)

Coefficient de corrélation de Pearson

Pour chez vous

Calcul manuel :

N = 9

$\bar{X} = 3.89$

$\bar{Y} = 27.33$

$\Sigma X_i^2 = 169$

$\Sigma Y_i^2 = 7104$

$\Sigma X_i Y_i = 1058$

X	Y	X^2	Y^2	XY
1	20	1	400	20
2	18	4	324	36
2	25	4	625	50
3	27	9	729	81
4	24	16	576	96
5	28	25	784	140
5	29	25	841	145
6	35	36	1225	210
7	40	49	1600	280
$\bar{X} = 3.89$	$\bar{Y} = 27.33$	Σ= 169	Σ= 7104	Σ= 1058

$$r(X,Y) = \frac{\sum(XiYi) - N\bar{X}\bar{Y}}{\sqrt{(\sum X_i^2 - N\bar{X}^2) * (\sum Y_i^2 - N\bar{Y}^2)}}$$

$$r = \frac{1058 - (9 * 3.89 * 27.33)}{\sqrt{(169 - 9 * 3.89^2) * (7104 - 9 * 27.33^2)}} = \frac{101.33}{111.79} = \mathbf{0.90}$$

Coefficient de corrélation de Pearson

Pour chez vous

Table Pearson (bilatérale) pour ddl = *n*-2 = 7 et α=0,05

$r_{théo} = 0.6664$

$|r_{obs}| = 0.90$

Table de valeurs critiques $r_{\alpha/2}$ pour effectuer un test bilatéral pour un ddl $\upsilon = n - 2$

ν \ P	0,05	0,01	0,001
1	0,9969	0,9999	1,0000
2	0,9500	0,9900	0,9990
3	0,8783	0,9587	0,9912
4	0,8114	0,9172	0,9741
5	0,7545	0,8745	0,9507
6	0,7067	0,8343	0,9249
7	0,6664	0,7977	0,8982
8	0,6319	0,7646	0,8721
9	0,6021	0,7348	0,8471
10	0,5760	0,7079	0,8233
11	0,5529	0,6835	0,8010
12	0,5324	0,6614	0,7800

Règle de décision:

$|r_{obs}| > r_{théo}$: H0 est rejetée au profit de H1.

→ r est significativement différent de 0.

Il y existe une relation linéaire positive entre les 2 variables, elles sont positivement et fortement corrélées (r = 0.90).

Partie 5. Etude des relations entre variables

5.1. Le coefficient de corrélation

5.1.1. Coefficient de corrélation de Pearson

5.1.2. Coefficient de corrélation de Spearman

5.2. De la corrélation au modèle de régression

5.2.1. Régression linéaire simple

5.2.2. Régression linéaire multiple

Coefficient de corrélation de Spearman

Effectuer ce test sous R

Code :

Préciser la méthode

```
cor.test(Forets$RS_Ligneux,Forets$RS_herbaces, method="spearman")
```

Sortie :

```
        Spearman's rank correlation rho

data:  Forets$RS_Ligneux and Forets$RS_herbaces
S = 11.089, p-value = 0.0007208
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9075951
```

Forte relation monotone positive entre les 2 variables (r = 0.91), hautement significative.

Coefficient de corrélation de Spearman

Pour chez vous

Calcul (pour info) :

1. Attribuer un rang à chaque valeur X_i du couple XY en fonction d'un ordre croissant
   → rangs des X_i (rang moyen en cas d'ex-aequos)
2. Faire la même chose avec les valeurs de Y → rangs des Y_i
3. Faire la différence entre les deux rangs attribués à chaque couple XY et l'élever au carré
   → $d = (X_i - Y_i)^2$
4. Faire la somme des différences élevées au carré = Σd^2
5. Appliquer la formule suivante :

$$r_s = 1 - \left(\frac{6 \sum d^2}{N(N^2-1)}\right)$$ avec N = nombre de couples XY

Coefficient de corrélation de Spearman

Pour chez vous

Test de significativité :

Hypothèses:

H_0 : $r_s = 0$, les rangs de X et Y sont indépendantes

H_1 : $r_s \neq 0$, les rangs des deux variables ne sont pas indépendants

- **Si N ≥ 30**, comparer $r_s \times \sqrt{N-1}$ aux valeurs de la LNCR
- **Si N < 30**, on compare r_s avec des valeurs extraites de la table de Spearman (bilatéral, pour α = 0.05 et ddl = N-2)

Règle de décision:

Si $|r_{S_obs}| > r_{S_théo}$ → H_1 : le coefficient est jugé significatif

→ corrélation/liaison monotone significative

Table H. Critical Values of r_s (Spearman Rank-Order Correlation Coefficient)

	Level of Significance for One-Tailed Test			
	.05	.025	.01	.005
	Level of Significance for Two-Tailed Test			
df	.10	.05	.02	.01
5	.900	1.000	1.000	—
6	.829	.886	.943	1.000
7	.714	.786	.893	.929
8	.643	.738	.833	.881
9	.600	.683	.783	.833
10	.564	.648	.746	.794
12	.506	.591	.712	.777
14	.456	.544	.645	.715
16	.425	.506	.601	.665
18	.399	.475	.564	.625
20	.377	.450	.534	.591
22	.359	.428	.508	.562
24	.343	.409	.485	.537
26	.329	.392	.465	.515
28	.317	.377	.448	.496
30	.306	.364	.432	.478

Coefficient de corrélation de Spearman

Pour chez vous

Calcul manuel :

Lgnx	Herb	Rangs_Lgnx	Rangs_herb	d = Rangs(L-H)	d^2
1	20	1	2	1	1
2	18	2.5	1	1.5	2.25
2	25	2.5	4	1.5	2.25
3	27	4	5	1	1
4	24	5	3	2	4
5	28	6.5	6	0.5	0.25
5	29	6.5	7	0.5	0.25
6	35	8	8	0	0
7	40	9	9	0	0

$\Sigma d^2 = 11$

$$r_s = 1 - \left(\frac{6\sum d^2}{n(n^2-1)}\right) = 1 - \frac{6\times 11}{9(9^2-1)} = 1 - \frac{66}{720} = 0.9084$$

Coefficient de corrélation de Spearman

Pour chez vous

N < 30 → Table de Spearman (bilatéral, pour α = 0.05 et ddl = N-2 = 7).

$r_{S_théo} = 0.786$

$|r_{S_obs}| = 0.9084$

Règle de décision:

$|r_{S_obs}| > r_{S_théo}$: H0 rejetée, on accepte H1

→ r_S est significativement différent de 0.

Table H. Critical Values of r_s (Spearman Rank-Order Correlation Coefficient)

	Level of Significance for One-Tailed Test			
	.05	.025	.01	.005
	Level of Significance for Two-Tailed Test			
df	.10	.05	.02	.01
5	.900	1.000	1.000	—
6	.829	.886	.943	1.000
7	.714	.786	.893	.929
8	.643	.738	.833	.881
9	.600	.683	.783	.833
10	.564	.648	.746	.794
12	.506	.591	.712	.777
14	.456	.544	.645	.715
16	.425	.506	.601	.665
18	.399	.475	.564	.625
20	.377	.450	.534	.591
22	.359	.428	.508	.562
24	.343	.409	.485	.537
26	.329	.392	.465	.515
28	.317	.377	.448	.496
30	.306	.364	.432	.478

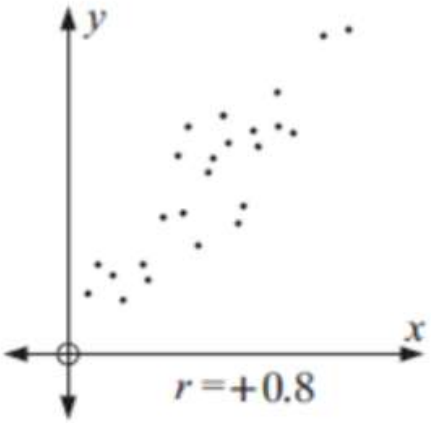
Il y existe une relation positive monotone croissante significative entre les 2 variables, elles sont positivement et fortement corrélées (r=0.91, α=5%).

ATTENTION➔ Corrélation ne signifie pas causalité

Corrélation: lien statistique, sans qu'on se demande quelle variable agit sur l'autre (rôle symétrique de X <--> Y).

Causalité: lien qui affirme qu'une variable agit sur une autre → nécessite une expérience planifiée dans le but de répondre à CETTE question (contrôle des facteurs secondaires)!

Ex: On note chaque semaine pendant 1 an le nombre de cornets de glace à la vanille consommés et nombre de d'hospitalisation de personnes âgées.



Consommation de cornets de glace à la vanille

Hospitalisation des personnes âgées

Eté: Fortes températures (canicule)

= Facteurs de confusion





Le paradoxe de Simpson — Science étonnante #7

Facteur de confusion et paradoxe de Simpson :

https://www.youtube.com/watch?v=vs_Zzf_vL2I