Ecole Supérieure en Sciences Appliquées de Tlemcen

Analyse de données

Chapitre 3: Description bidimensionnelle et mesure de liaison entre variables

Partie 1

Présentée par:

Dr Imane NEDJAR

Introduction

L’objectif de la Statistique Descriptive est de décrire les données observées pour mieux les analyser

- **Liaison Entre Variables**
- **Régression linéaire**

Liaison Entre Variables

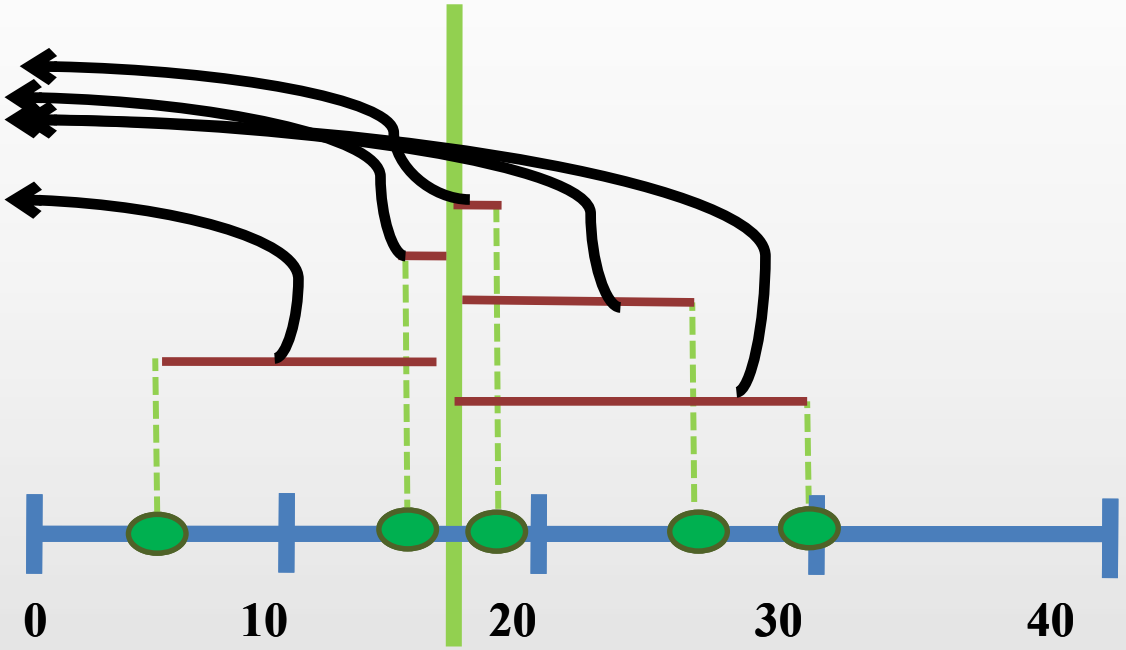
Covariance

Variance

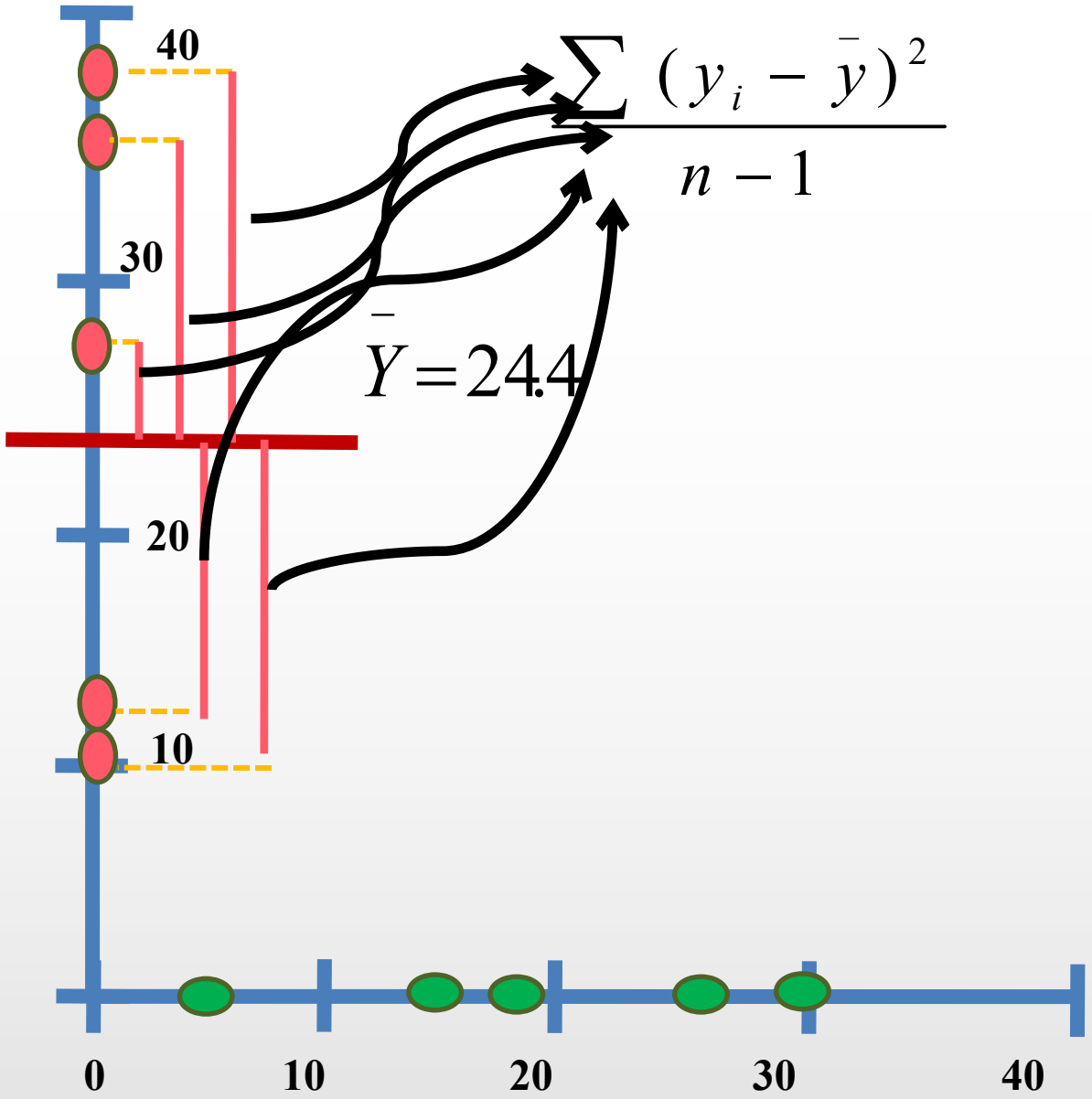
$$\frac{\sum (x_i - \bar{x})^2}{n-1}$$

Moyenne

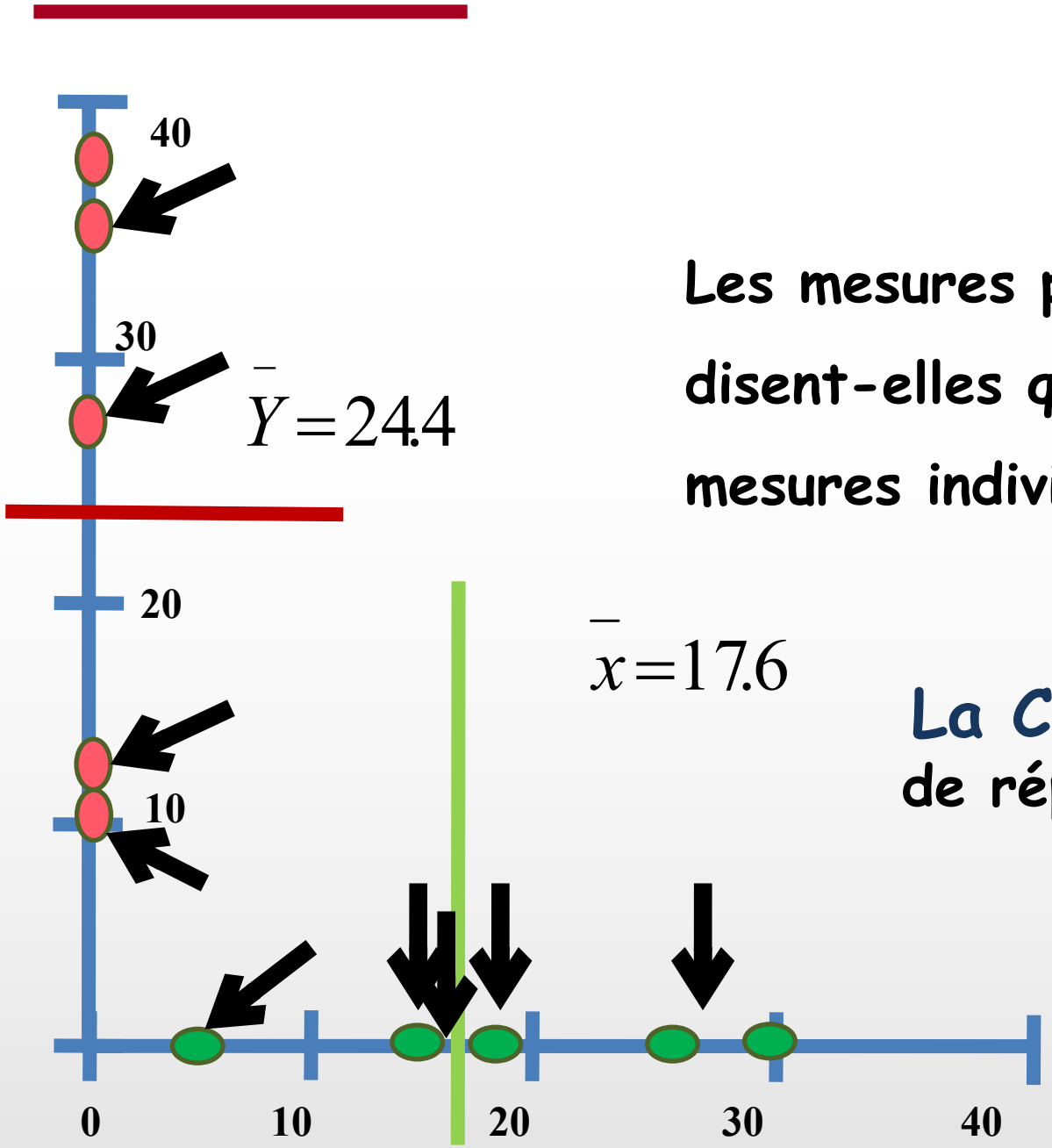
$$\bar{x} = 17.6$$



Covariance



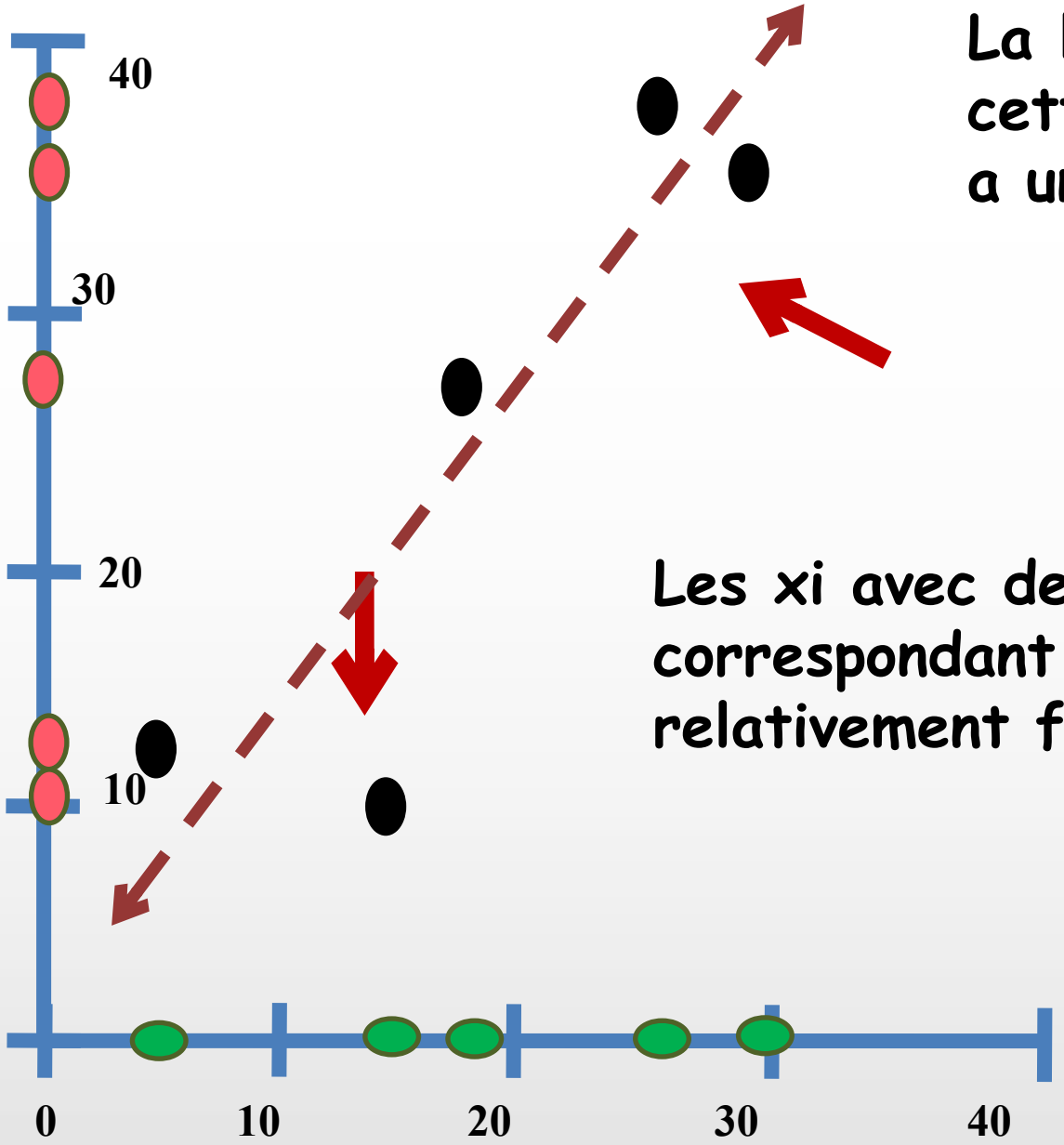
Covariance



Les mesures prises par paires nous disent-elles quelque chose que les mesures individuelles ne disent pas ?

La Covariance est une façon de répondre à cette question

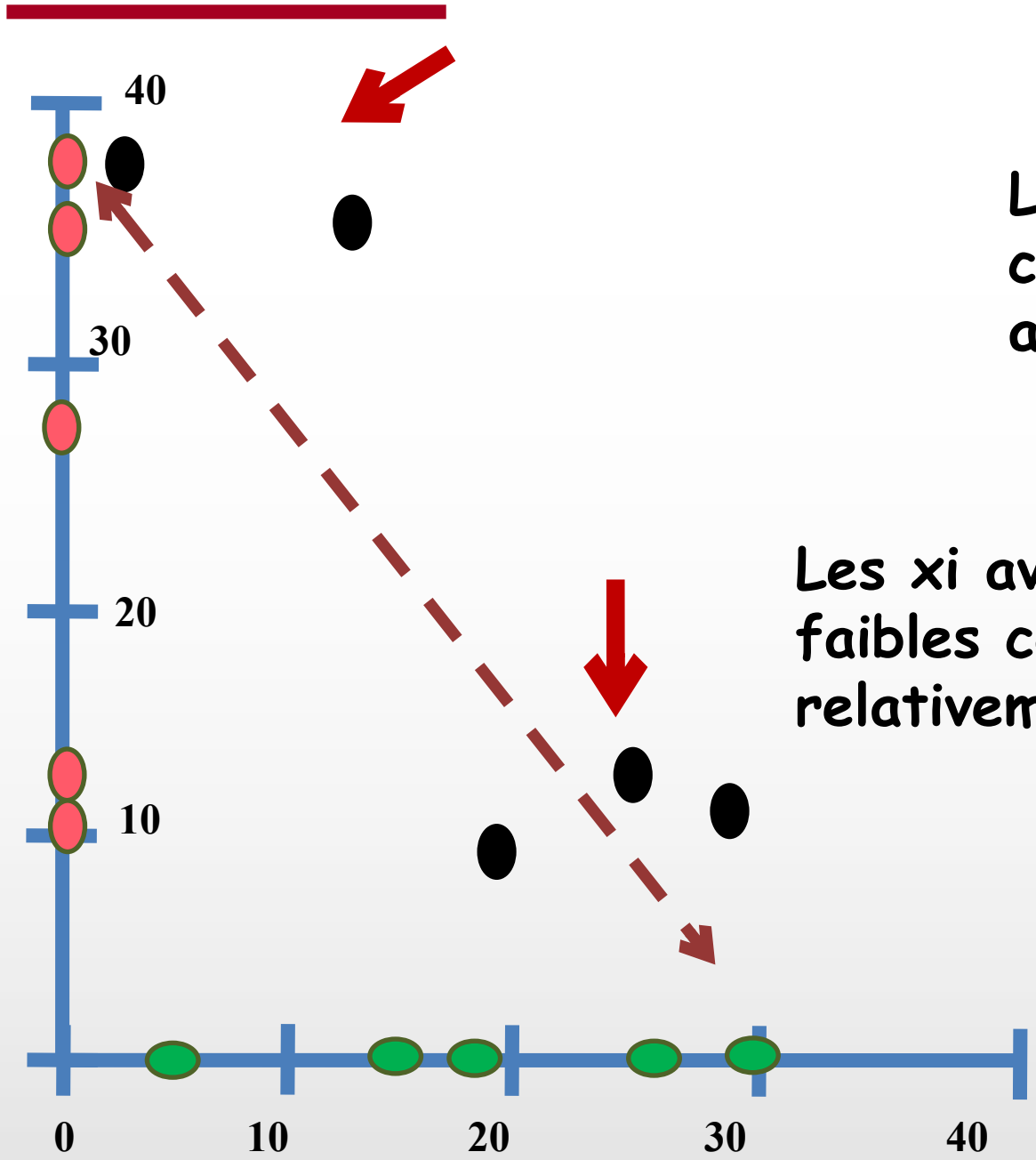
Covariance



La ligne qui représente cette relation particulière a une pente positive

Les xi avec des valeurs faibles correspondant également à des valeurs relativement faibles pour yi

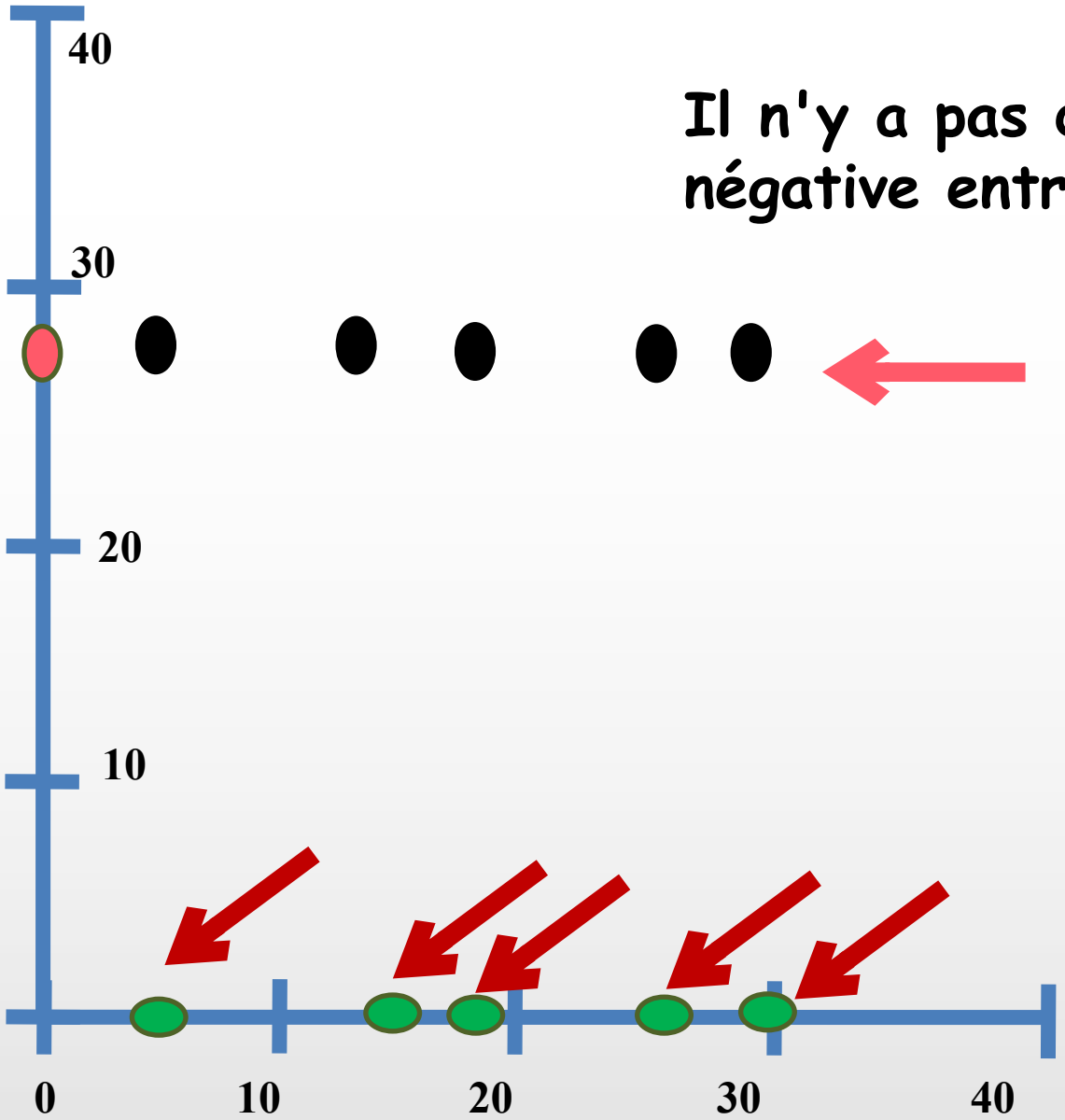
Covariance



La ligne qui représente cette relation particulière a une pente négative

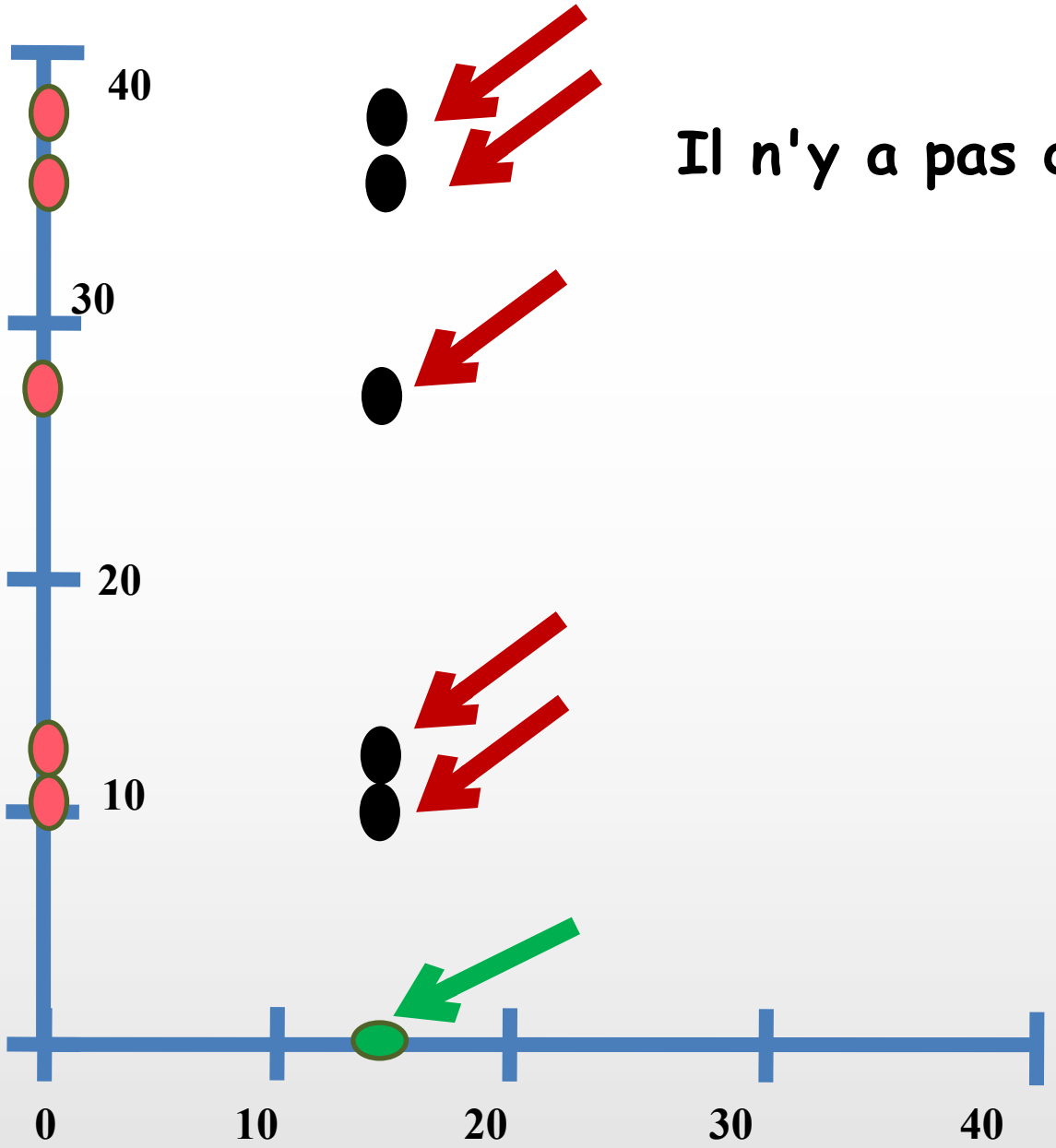
Les xi avec des valeurs relativement faibles correspondant à des valeurs relativement grandes pour yi

Covariance



Il n'y a pas de relation positive ou négative entre xi et yi

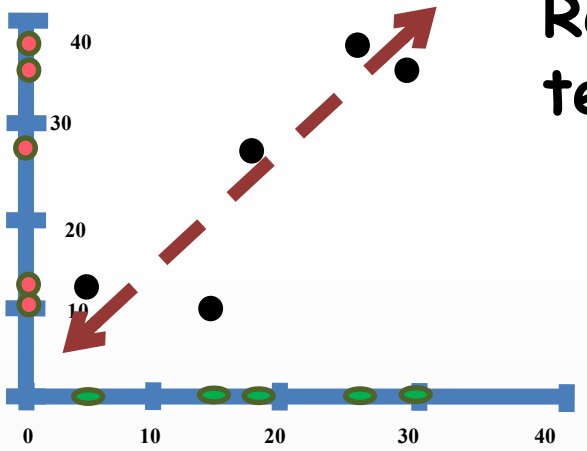
Covariance



Il n'y a pas de relation entre xi et yi

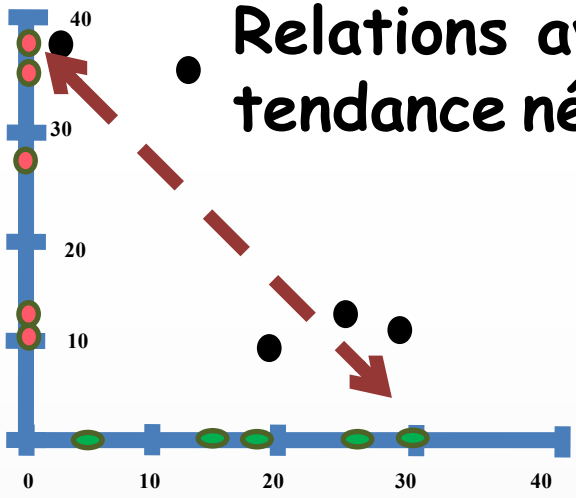
Covariance

La covariance peut classer ces trois types de relations

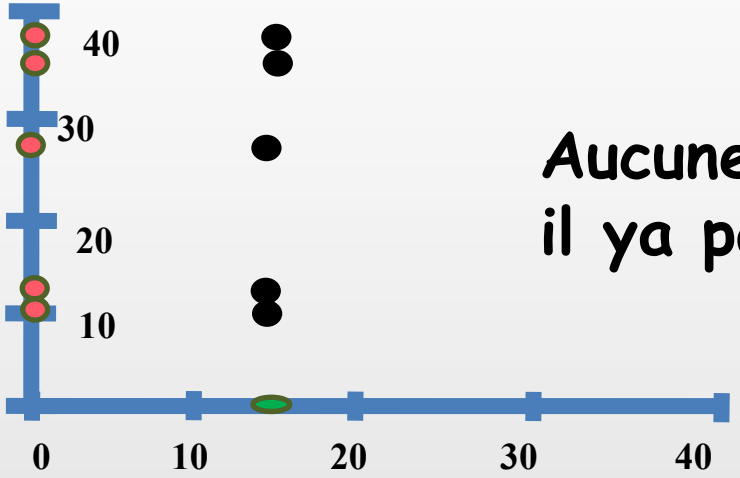


Relations avec une tendance positive

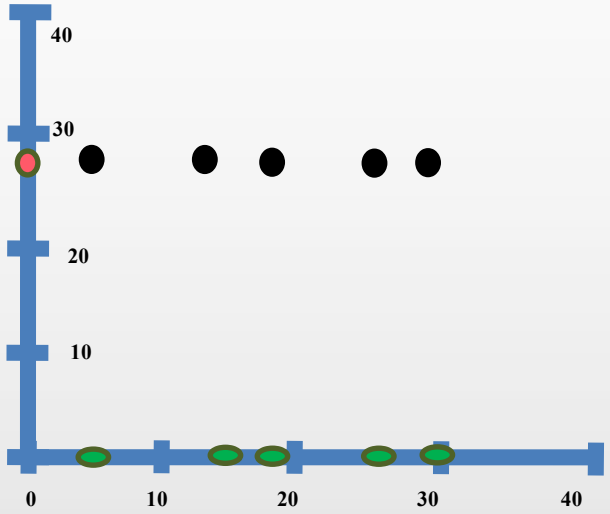
$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



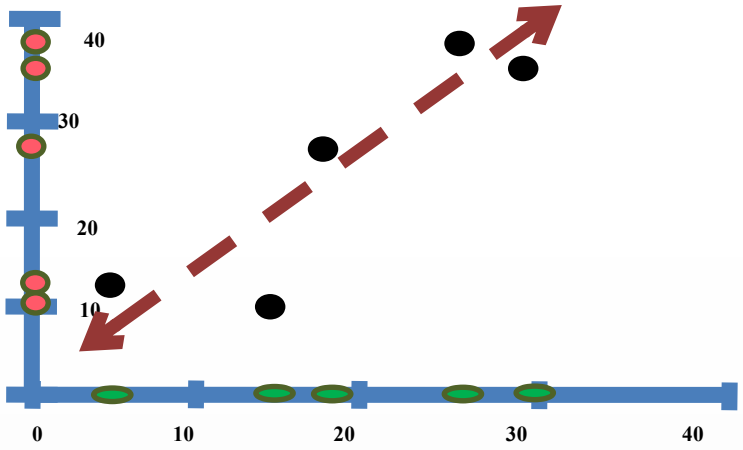
Relations avec une tendance négative

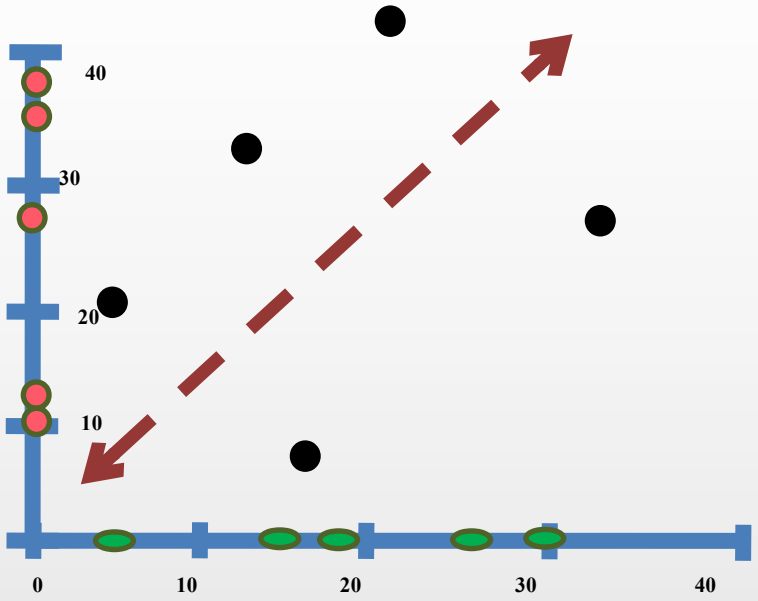


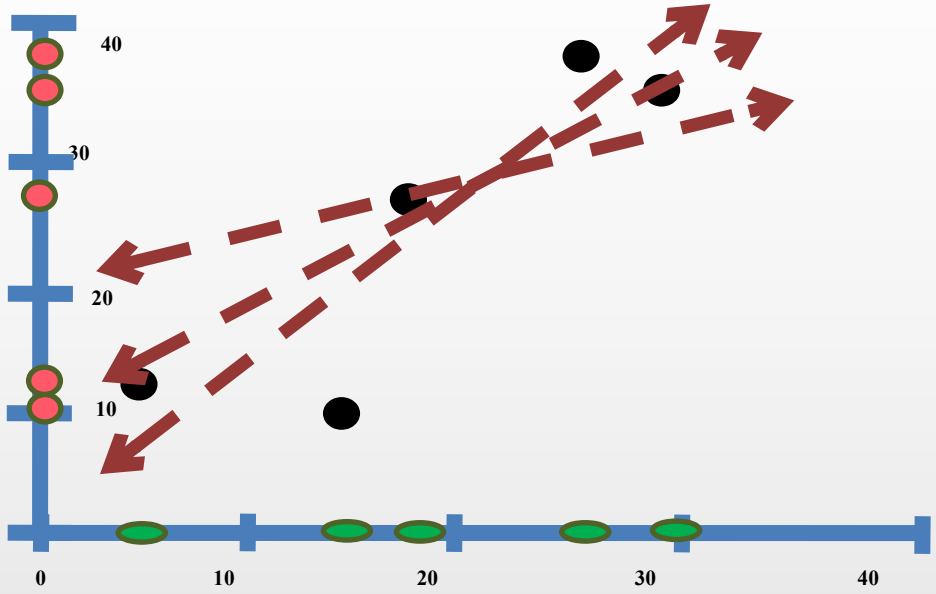
Aucune relations parce que il ya pas une tendance



Covariance


40
30
20
10
0
10
20
30
40


40
30
20
10
0
10
20
30
40


40
30
20
10
0
10
20
30
40

Correlation

La covariance est un tremplin de calcul vers quelque chose d'intéressant comme la corrélation

$$corr\ (X, Y) = \frac{Cov\ (X, Y)}{\sigma x\, \sigma y}$$

➔ $-1 \leq \text{corr}_{XY} \leq +1$

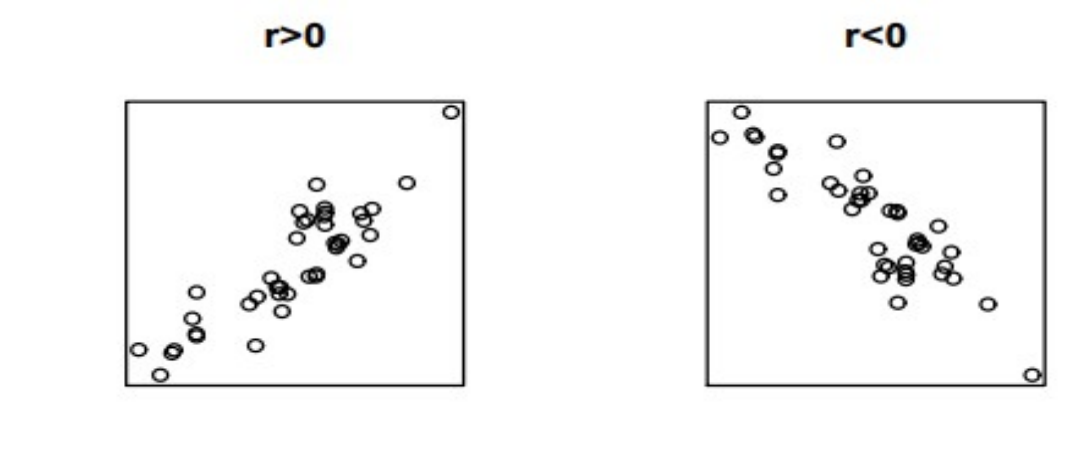
➔ La valeur absolue du coefficient indique l'intensité de la liaison. $0 \leq |\ \text{corr}_{xy}\ | \leq 1$

➔ Le coefficient de corrélation est symétrique : $\text{corr}_{XY} = \text{corr}_{YX}$
cette propriété est évidente au vu de la définition de corr_{XY}

Correlation

➔ Le signe du coefficient indique le sens de la liaison :

- ➢Si corr est positif, les points sont alignés le long d'une droite croissante= les deux variables ont tendance à varier dans le même sens.
- ➢Si corr est négatif, les points sont alignés le long d'une droite décroissante= les deux variables ont tendance a` varier en sens opposés



Liaison entre

Variable quantiative

Le coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson
mesure la force et la direction d'une relation linéaire entre deux variables

- **Pearson renvoie une valeur comprise entre -1 et 1**
- **+1 une corrélation positive parfaite**
- **-1 une corrélation négative parfaite entre les rangs**
- **0 = aucune corrélation entre les rangs.**

Le coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson
mesure la force et la direction d'une relation linéaire entre deux variables

$$r = \frac{n\sum x_i y_i - \sum x_i \times \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \times \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Le coefficient de corrélation de Pearson

X	Y
1	10
2	20
3	30
4	40
5	50

X	Y	X^2	Y^2	XY
1	10	1	100	10
2	20	4	400	40
3	30	9	900	90
4	40	16	1600	160
5	50	25	2500	250
$\sum X=15$	**$\sum Y=150$**	**$\sum X^2=55$**	**$\sum Y^2=5500$**	**$\sum XY=550$**

$$r = \frac{5\times 550 - 15\times 150}{\sqrt{5\times 55-(15)^2}\times\sqrt{5\times 5500-(150)^2}} = 1$$

Liaison entre

Variable ordinales

Le coefficient de correlation de Spearman (Rho)

Le coefficient de corrélation τ Kendall

Le coefficient de Spearman (Rho)

Le coefficient de corrélation de Spearman mesure la force et la direction d'une association entre deux variables ordinaire qui ne sont pas fortement linéaire

- **Spearman renvoie une valeur comprise entre -1 et 1**
- **+1 une corrélation positive parfaite**
- **-1 une corrélation négative parfaite entre les rangs**
- **0 = aucune corrélation entre les rangs.**

Le coefficient de Spearman (Rho)

L'idée est de substituer aux valeurs observées leurs rangs. Par la création des nouvelles colonnes

Ri = Rang(xi): correspond au rang i de l'observation xi dans la colonne des X ;

Si = Rang(Yi): correspond au rang i de l'observation yi dans la colonne des Y ;

Di = Ri − Si

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}$$

Le coefficient de Spearman (Rho)

Exemple

Individu	Concours 1	Concours 2
S1	7	10
S2	10	12
S3	1	4
S4	6	7
S5	9	11
S6	13	9
S7	3	2
S8	5	4
S9	11	5
S10	9	11
S11	6	6
S12	4	1

Individu	Concours 1
S1	7
S2	10
S3	1
S4	6
S5	9
S6	13
S7	3
S8	5
S9	11
S10	9
S11	6
S12	4

Individu	Concours 2
S1	10
S2	12
S3	4
S4	7
S5	11
S6	9
S7	2
S8	4
S9	5
S10	11
S11	6
S12	1

Le coefficient de Spearman (Rho)

Exemple

Individu	Concours 1
S1	7
S2	10
S3	1
S4	6
S5	9
S6	13
S7	3
S8	5
S9	11
S10	9
S11	6
S12	4

1. *On ordonne les individus selon chaque variable*
2. **On affecte un rang brut**
3. **Les Ex aequo on calcule leur moyen**

Individu	Concours1	Rangs bruts	Rangs
S3	1	1	1
S7	3	2	2
S12	4	3	3
S8	5	4	4
S4	6	5	5.5
S11	6	6	5.5
S1	7	7	7
S5	9	8	8.5
S10	9	9	8.5
S2	10	10	10
S9	11	11	11
S6	13	12	12

Le coefficient de Spearman (Rho)

Exemple

Individu	Concours 2
S1	10
S2	12
S3	4
S4	7
S5	11
S6	9
S7	2
S8	4
S9	5
S10	11
S11	6
S12	1

1. *On ordonne les individus selon chaque variable*
2. **On affecte un rang brut**
3. **Les Ex aequo on calcule leur moyen**

Individu	Concours2	Rangs bruts	Rangs
S12	1	1	1
S7	2	2	2
S8	4	3	3.5
S3	4	4	3.5
S9	5	5	5
S11	6	6	6
S4	7	7	7
S6	9	8	8
S1	10	9	9
S5	11	10	10.5
S10	11	11	10.5
S2	12	12	12

Le coefficient de Spearman (Rho)

Exemple

Individus	Concours1	Concours2	d	D*D
S1	7	9	-2	4
S2	10	12	-2	4
S3	1	3.5	-2.5	6.25
S4	5.5	7	-1.5	2.25
S5	8.5	10.5	-2	4
S6	12	8	-4	16
S7	2	2	0	0
S8	4	3.5	-0.5	0.25
S9	11	5	6	36
S10	8.5	10.5	-2	4
S11	5.5	6	-0.5	0.25
S12	3	1	2	4
Total				**81**

$$\rho = 1 - \frac{6*81}{12(12^2 - 1)} = 0.72$$

τ Kendall

C'est un coefficient qui représente le degré de concordance entre deux variables

- **τ Kendall renvoie une valeur comprise entre -1 et 1**
- **+1 une corrélation positive parfaite**
- **-1 une corrélation négative parfaite entre les rangs**
- **0 = aucune corrélation entre les rangs.**

τ Kendall

On commence par trier les observations par ordre croissant. Cela nous donne donc un ordre parfait sur la première variable mais pas sur la deuxième variable. On va alors comparer toutes les paires possibles.

- **Paire concordantes (Nc)**

Le nombre des exemples observés en dessous d'un exemple particulier et qui sont plus grands que ce exemple particulier

- **Paire discordante (Nd)**

➢Le nombre des exemples observés en dessous d'un exemple particulier et qui sont plus petits que ce exemple particulier

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n (n - 1)}$$

τ Kendall

Exemple

Sujet	V1	V2
1	1,1	3,4
2	5	6,5
3	2,4	2,8
4	3	1
5	3,2	2
6	2	1,1

Trie selon V1

Sujet	V1	V2	Nc (+1)/Nd (-1)				
1	1,1	3,4	_	_	_	_	_
6	2	1,1	-1	_	_	_	_
3	2,4	2,8	-1	+1	_	_	_
4	3	1	-1	-1	-1	_	_
5	3,2	2	-1	+1	-1	+1	_
2	5	6,5	+1	+1	+1	+1	+1

$$\tau = \frac{8 - 7}{\frac{1}{2} 6 (6 - 1)} = 0{,}067$$

Liaison entre

Variable nominales

Khi-deux, ou Chi-2, ou encore $\chi 2$

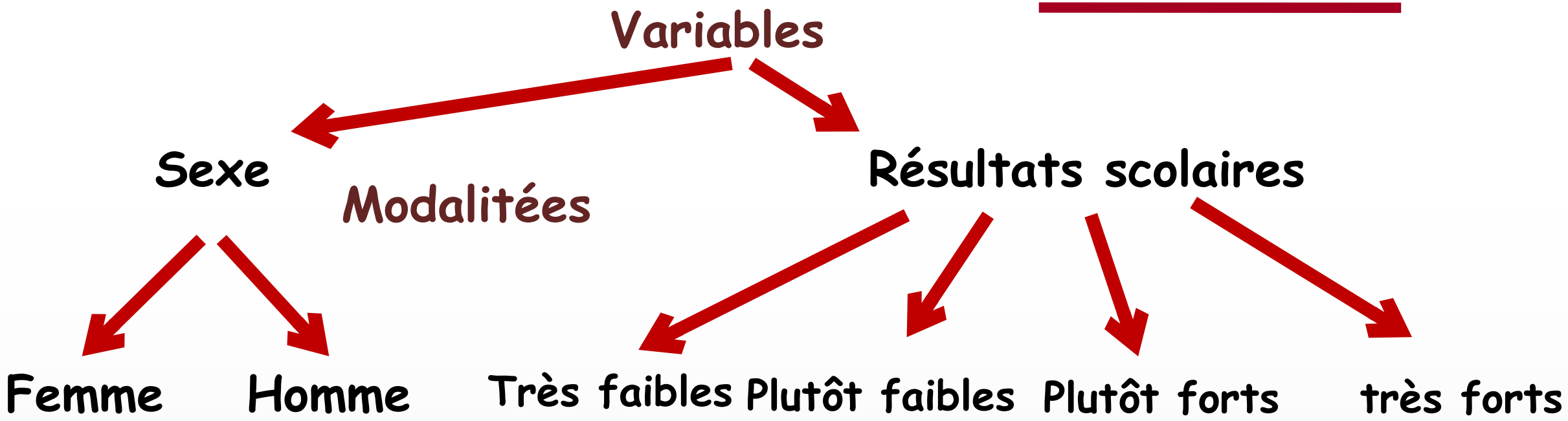
Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore $\chi 2$)

Le test d'indépendance entre deux variables , ou **test du khi** deux donne la possibilité de vérifier si les données provenant d'un échantillon aléatoire permettent de conclure à **l'indépendance** entre deux **variable qualitatives** dans la population d'ou a été trié cet échantillon

Exemple

On va vérifier s'il ya un lien entre **le sexe** et **les résultats scolaires** , on pourrait alors choisir de noter, chez un certain nombre d'élèves, s'ils sont de **sexe féminin** ou **masculin** et si leurs résultats scolaires **sont très faibles**, **plutôt faibles**, **plutôt forts** ou **très forts**.

Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore $\chi 2$)



	Femme	Homme	Total
Très faibles	8	20	28
Plutôt faibles	14	45	59
Plutôt forts	32	31	63
Très forts	30	20	50
Total	84	116	200

Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore $\chi 2$)

Exemple

L'hypothèse nulle (H0):

« Le sexe et les résultats scolaires sont indépendants »

La contre hypothèse (H1):

« Il ya un lien entre les résultats scolaires et le sexe »

- L'hypothèse nulle H0 est dite prudente
- Elle est considérée comme **vraie** tout au long du test et on ne la rejettera que si on a suffisamment de preuves contre elle .

Exemple

S'il n'ya **aucun** rapport entre les deux variable, les effectifs doivent être repartis comme suite

Tableau effectif théorique

	Femme	Homme	Total
Très faibles	11.76	16.24	28
Plutôt faibles			59
Plutôt forts			63
Très forts			50
Total	84 **(42%)**	116 **(58%)**	200

28 élèves ont des résultats qui sont très faibles , **42%** parmi eux doivent être des filles , alors le nombre de fille ayant des résultats très faibles est **28*42%=11.76**

Exemple

S'il n'ya **aucun** rapport entre les deux variable, les effectifs doivent être repartis comme suite

Tableau effectif théorique

	Femme	Homme	Total
Très faibles			28
Plutôt faibles			59
Plutôt forts			63
Très forts			50
Total	84 **(42%)**	116 **(58%)**	200

28 élèves ont des résultats qui sont très faibles , **42%** parmi eux doivent être des filles

Exemple

S'il n'ya **aucun** rapport entre les deux variable, les effectifs doivent être repartis comme suite

Tableau effectif théorique

	Femme	Homme	Total
Très faibles	11.76	16.24	28
Plutôt faibles	24.78	34.22	59
Plutôt forts	26.46	36.54	63
Très forts	21.00	29.00	50
Total	84 **(42%)**	116 **(58%)**	200

De telles fréquences , calculées selon **L'hypothèse** qu'il n'y a **aucun rapport** entre le sexe et les résultats scolaires , sont appelées fréquence théoriques et notées f_t

Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore $\chi 2$)

Exemple

Pour étudier le rapport entre les deux variables résultats scolaires et le sexe des élèves il faut

alors comparer **les fréquences théoriques f_t** et les **fréquences réellement** obtenues **f_o** (Observées)

Tableau effectif théorique

	Femme	Homme	Total
Très faibles	11.76	16.24	28
Plutôt faibles	24.78	34.22	59
Plutôt forts	24.46	36.54	63
Très forts	21.00	29.00	50
Total	84 **(42%)**	116 **(58%)**	200

Tableau effectif réel

	Femme	Homme	Total
Très faibles	8	20	28
Plutôt faibles	14	45	59
Plutôt forts	32	31	63
Très forts	30	20	50
Total	84	116	200

Si les variables sexe et résultats scolaires étaient parfaitement indépendants alors $f_o = f_t$

Exemple

On remarque que $\mathbf{f_t} \neq \mathbf{f_o}$

Il existe un certain lien entre les deux variables

Il faut mesurer ce lien

=> Pour ce faire on calcule le carré de contingence khi-2 $\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{c}\frac{(o_{ij} - t_{ij})^2}{t_{ij}}$

Tableau effectif théorique

	Femme	Homme	Total
Très faibles	11.76	16.24	28
Plutôt faibles	24.78	34.22	59
Plutôt forts	24.46	36.54	63
Très forts	21.00	29.00	50
Total	84 **(42%)**	116 **(58%)**	200

Tableau effectif réel

	Femme	Homme	Total
Très faibles	8	20	28
Plutôt faibles	14	45	59
Plutôt forts	32	31	63
Très forts	30	20	50
Total	84	116	200

Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore χ2)

Exemple

Tableau effectif réel

	Femme	Homme	Total
Très faibles	8	20	28
Plutôt faibles	14	45	59
Plutôt forts	32	31	63
Très forts	30	20	50
Total	84	116	200

Tableau effectif théorique

	Femme	Homme	Total
Très faibles	11.76	16.24	28
Plutôt faibles	24.78	34.22	59
Plutôt forts	24.46	36.54	63
Très forts	21.00	29.00	50
Total	84 **(42%)**	116 **(58%)**	200

$$\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{c}\frac{(o_{ij} - t_{ij})^2}{t_{ij}}$$

$$\chi^2 = \frac{(8-11.76)^2}{11.76} + \frac{(20-16.24)^2}{16.24} + \frac{(14-24.78)^2}{24.78} + \frac{(45-34.22)^2}{34.22} + \frac{(32-24.46)^2}{24.46} + \frac{(31-36.54)^2}{36.54} + \frac{(30-21.00)^2}{21.00} + \frac{(20-29)^2}{29} = 18.809$$

Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore $\chi 2$)

- Il faut noter que l'utilisation du χ^2 est déconseillée lorsque l'effectif théorique de certaines cases est petit (plus petit que 5). Il faut donc disposer d'autres indices.
- Plus la valeur du $\chi 2$ est grande, plus de degré d'association entre les deux variables est grand.
- S'il n'existe aucun lien entre deux variables ($f_o = f_t$) $\chi 2 = 0$

Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore χ2)

On fixe le seuil de signification

Le seuil de signification α , représente la probabilité du risque d'erreur qu'on est prêt a assumer en rejetant une hypothèse nulle

$$\alpha = 1\% \ , \ \alpha = 5\% \ , \ \alpha = 10\%$$

Le degré de liberté V(nu)

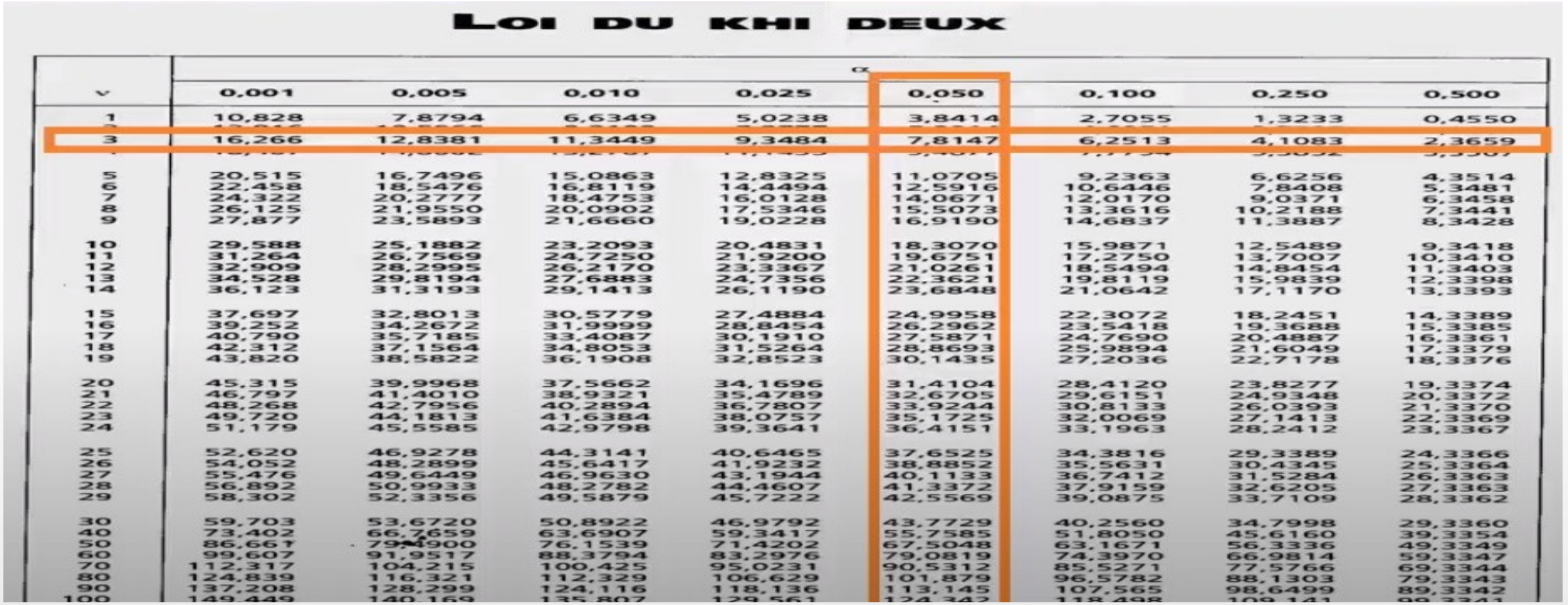
$$v=(\text{nombre de lignes}-1)*(\text{nombre de colonnes}-1)$$

$$v=(4-1)*(2-1)=3$$

	Femme	Homme	Total
Très faibles	8	20	28
Plutôt faibles	14	45	59
Plutôt forts	32	31	63
Très forts	30	20	50
Total	84	116	200

Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore $\chi 2$)

Table présente les valeur critiques du carré de contingence pour différents seuils de signification de α et pour différent valeur de v



LOI DU KHI DEUX

v	α 0,001	0,005	0,010	0,025	0,050	0,100	0,250	0,500
1	10,828	7,8794	6,6349	5,0238	3,8414	2,7055	1,3233	0,4550
3	16,266	12,8381	11,3449	9,3484	7,8147	6,2513	4,1083	2,3659
5	20,515	16,7496	15,0863	12,8325	11,0705	9,2363	6,6256	4,3514
6	22,458	18,5476	16,8119	14,4494	12,5916	10,6446	7,8408	5,3481
7	24,322	20,2777	18,4753	16,0128	14,0671	12,0170	9,0371	6,3458
8	26,125	21,9550	20,0902	17,5346	15,5073	13,3616	10,2188	7,3441
9	27,877	23,5893	21,6660	19,0228	16,9190	14,6837	11,3887	8,3428
10	29,588	25,1882	23,2093	20,4831	18,3070	15,9871	12,5489	9,3418
11	31,264	26,7569	24,7250	21,9200	19,6751	17,2750	13,7007	10,3410
12	32,909	28,2995	26,2170	23,3367	21,0261	18,5494	14,8454	11,3403
13	34,528	29,8194	27,6883	24,7356	22,3621	19,8119	15,9839	12,3398
14	36,123	31,3193	29,1413	26,1190	23,6848	21,0642	17,1170	13,3393
15	37,697	32,8013	30,5779	27,4884	24,9958	22,3072	18,2451	14,3389
16	39,252	34,2672	31,9999	28,8454	26,2962	23,5418	19,3688	15,3385
17	40,790	35,7185	33,4087	30,1910	27,5871	24,7690	20,4887	16,3361
18	42,312	37,1564	34,8053	31,5264	28,8693	25,9894	21,6049	17,3379
19	43,820	38,5822	36,1908	32,8523	30,1435	27,2036	22,7178	18,3376
20	45,315	39,9968	37,5662	34,1696	31,4104	28,4120	23,8277	19,3374
21	46,797	41,4010	38,9321	35,4789	32,6705	29,6151	24,9348	20,3372
22	48,268	42,7956	40,2894	36,7807	33,9244	30,8133	26,0393	21,3370
23	49,720	44,1813	41,6384	38,0757	35,1725	32,0069	27,1413	22,3369
24	51,179	45,5585	42,9798	39,3641	36,4151	33,1963	28,2412	23,3367
25	52,620	46,9278	44,3141	40,6465	37,6525	34,3816	29,3389	24,3366
26	54,052	48,2899	45,6417	41,9232	38,8852	35,5631	30,4345	25,3364
27	55,476	49,6449	46,9630	43,1944	40,1133	36,7412	31,5284	26,3363
28	56,892	50,9933	48,2782	44,4607	41,3372	37,9159	32,6205	27,3363
29	58,302	52,3356	49,5879	45,7222	42,5569	39,0875	33,7109	28,3362
30	59,703	53,6720	50,8922	46,9792	43,7729	40,2560	34,7998	29,3360
40	73,402	66,7659	63,6907	59,3417	55,7585	51,8050	45,6160	39,3354
50	86,661	79,4900	76,1539	71,4202	67,5048	63,1671	56,3336	49,3349
60	99,607	91,9517	88,3794	83,2976	79,0819	74,3970	66,9814	59,3347
70	112,317	104,215	100,425	95,0231	90,5312	85,5271	77,5766	69,3344
80	124,839	116,321	112,329	106,629	101,879	96,5782	88,1303	79,3343
90	137,208	128,299	124,116	118,136	113,145	107,565	98,6499	89,3342
100	149,449	140,169	135,807	129,561	124,342	118,498	109,141	99,3341

$\alpha = 0.05$ v=3 => $\chi 2$ (critique)=7.8147

Le Khi-2 d'indépendance (ou Khi-deux, ou Chi-2, ou encore $\chi 2$)

$$\alpha = 0.05 \ v=3 \Rightarrow \chi 2 \text{ (critique)}=7.8147$$

$$\chi 2 = 18.809$$

On accepte ou on rejette l'hypothèse nulle ?

L'hypothèse nulle (H0):

« Le sexe et les résultats scolaires sont indépendants »

La contre hypothèse (H1):

« Il ya un lien entre les résultats scolaires et le sexe »

La règle de décision

- si $\chi 2 < \chi 2$ (critique) on accepte Ho
- sinon on rejette H0 et on accepte H1

On rejette l'hypothése H0 et on accepte H1 => Il existe un lien entre le sexe et les résultats scolaires

Liaison entre

Variable qualitative et une variable quantitative

Corrélation bisériale ponctuelle (Point Biserial Correlation)

Corrélation bisériale ponctuelle (Point Biserial Correlation)

Une estimation de la cohérence entre deux variables, dont l'une est dichotomique qualitative et l'autre quantitative

- **renvoie une valeur comprise entre -1 et 1**
- **+1 une corrélation positive parfaite**
- **-1 une corrélation négative parfaite entre les rangs**
- **0 = aucune corrélation entre les rangs.**

Corrélation bisériale ponctuelle (Point Biserial Correlation)

Une estimation de la cohérence entre deux variables, dont l'une est dichotomique qualitative et l'autre quantitative

$$r_{pq} = \frac{(\bar{x} - \bar{y})\sqrt{pq}}{sd}$$

Moyenne de la variable 1,2

Proportion dans la variable 1,2

Ecart type de la population

Corrélation bisériale ponctuelle (Point Biserial Correlation)

La relation significative existe entre le sexe du travailleur et l'exécution d'une tâche d'assemblage électronique

$$r_{pq} = \frac{(10 - 2.4)\sqrt{0.5 * 0.5}}{4.37} = 0.87$$

Sexe	Années
M	10
M	11
M	6
M	11
F	4
F	3
M	12
F	2
F	2
F	1

Sexe	Années
M	10
M	11
M	6
M	11
M	12
Means	10

Sexe	Années
F	4
F	3
F	2
F	2
F	1
Means	2.4