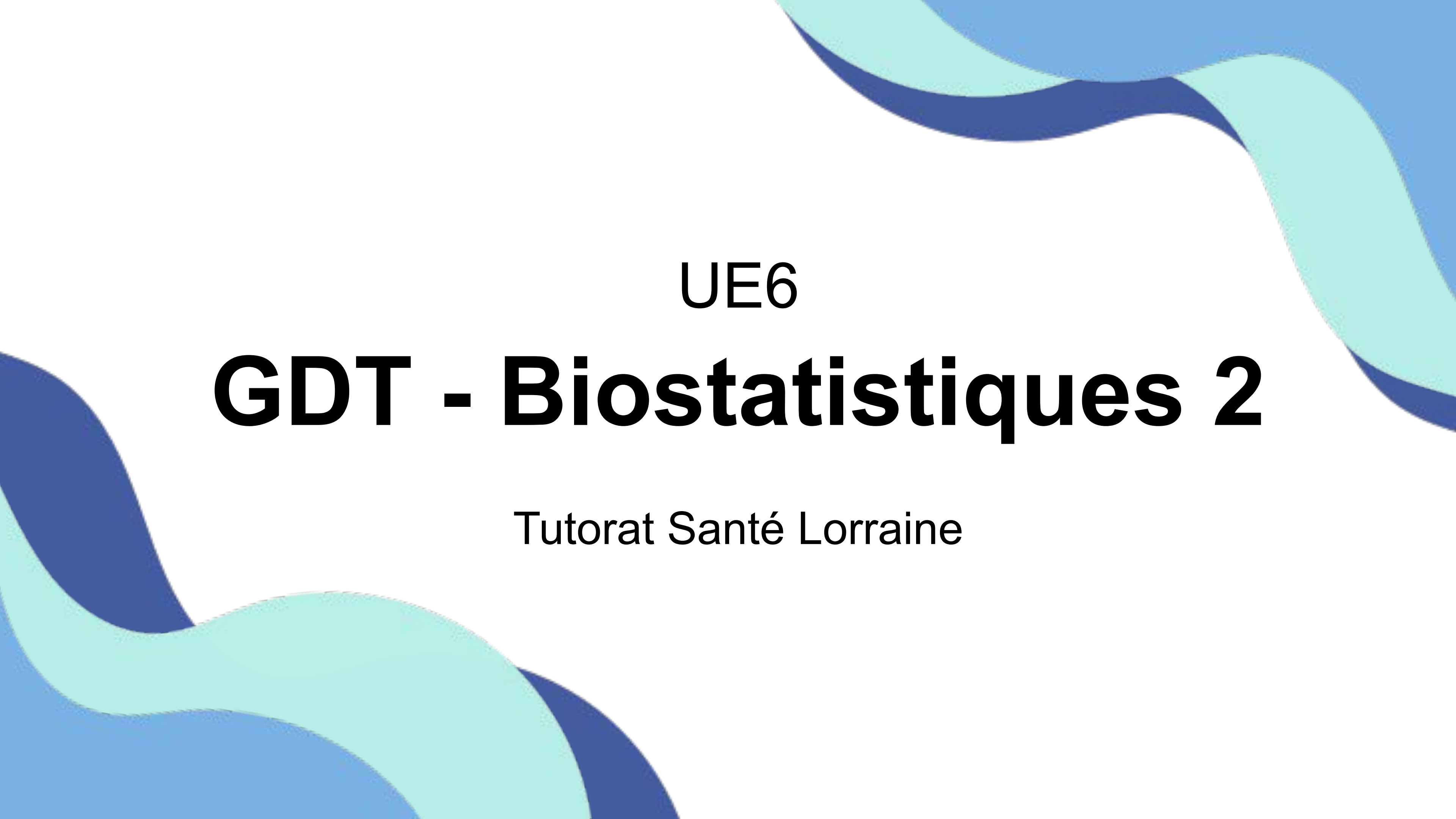




UE6

GDT - Biostatistiques 2

Tutorat Santé Lorraine

Introduction

Deuxième partie

Sommaire

1

**Régression et
corrélation**

2

**Tests
diagnostiques**

3

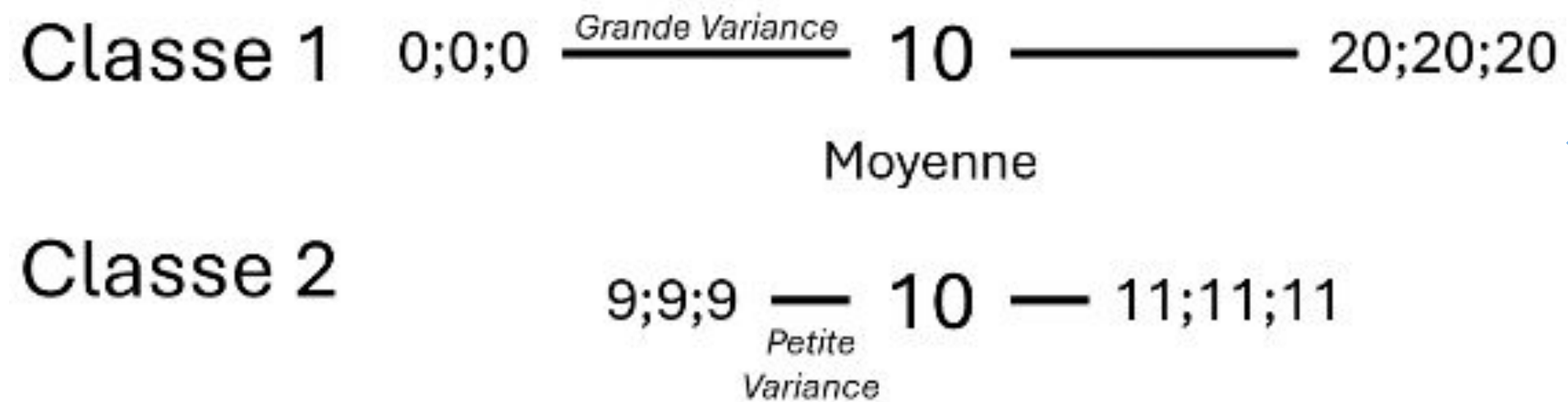
Courbe ROC

I - Régression et corrélation

Rappel

Moyenne : valeur centrale autour de laquelle les données se répartissent

Variance : mesure de la dispersion des valeur autour de la moyenne



2 classes avec une Moyenne de 10 $\neq$ répartition identique

I - Régression et corrélation

Rappel

Variance Générale : la variance correspond à la somme des carrés des écart à la moyenne

$$Var(X) = s_X^2 = \frac{SCE}{n} = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum (x_i^2)}{n} - \bar{x}^2$$

Variance de la population estimée : pour estimer la variance de la population on divise par “n-1”

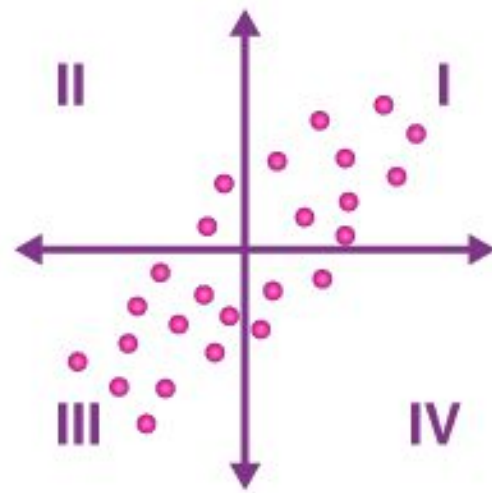
$$\frac{n}{n-1} s_X^2 = \frac{SCE}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{n}{n-1} \left(\frac{\sum (x_i^2)}{n} - \bar{x}^2 \right) = \frac{\sum (x_i^2) - n\bar{x}^2}{n-1}$$

I - Régression et corrélation

Rappel

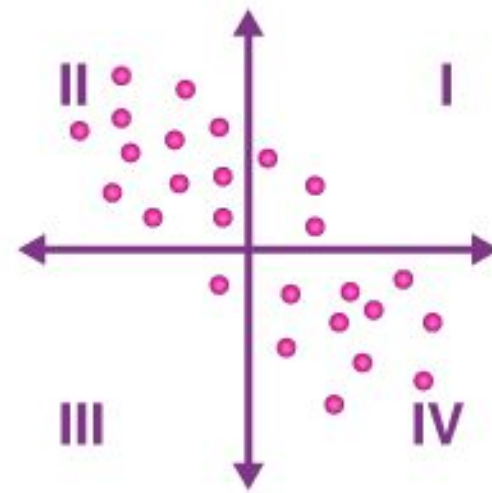
Covariance : quantifie la relation entre deux variables.

$cov > 0$



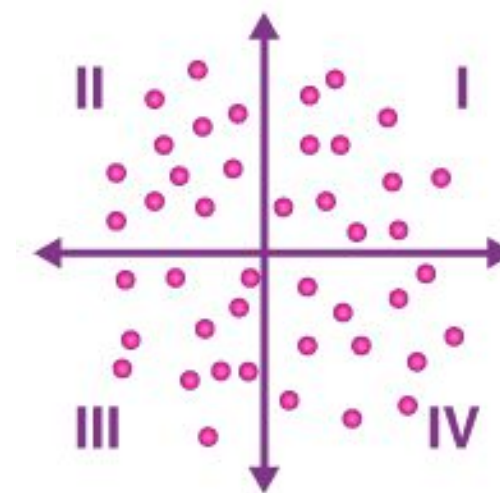
Liaison positive
 $Y \uparrow$ quand $X \uparrow$

$cov < 0$



Liaison négative
 $Y \downarrow$ quand $X \uparrow$

$cov = 0$



Pas de lien X et Y

I - Régression et corrélation

Rappel

Covariance globale : le calcul de la covariance est semblable à la variance, mais pour 2 variables

$$Cov(X, Y) = \frac{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})}{n} = \frac{\sum (x_i y_i)}{n} - \frac{\sum x_i \sum y_i}{n^2} = \frac{\sum (x_i y_i)}{n} - \bar{x} * \bar{y}$$

Covariance de la population estimée : on remplace n par “n-1”

$$\frac{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})}{n - 1} = \frac{n}{n - 1} \left(\frac{\sum (x_i y_i)}{n} - \frac{\sum x_i \sum y_i}{n^2} \right) = \frac{\sum (x_i y_i) - n * \bar{x} \bar{y}}{n - 1}$$

I - Régression et corrélation

Exercices

- A. Les variances et les covariances sont toujours positives.
- B. La covariance de deux variables X et Y est égale à la covariance de Y et X : $\text{cov}(X,Y) = \text{cov}(Y,X)$
- C. La variance d'une variable T est égale à la covariance de T avec elle même : $\text{var}(T) = \text{cov}(T,T)$
- D. La variance du triple d'une variable vaut le nonuple de la variance : $\text{var}(3T) = 9 \text{ var}(T)$
- E. La covariance entre une variable T et son opposée $-T$ est toujours négative

I - Régression et corrélation

Exercices

A. Les variances et les covariances sont toujours positives.

Les variances sont toujours positives

Les covariances peuvent être positive (relation +), négative (relation -) ou nulle (pas de relation)

B. La covariance de deux variables X et Y est égale à la covariance de Y et X : $cov(X,Y) = cov(Y,X)$

$$cov(X,Y) = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{n} = \frac{\sum(y_i - \bar{y})\sum(x_i - \bar{x})}{n} = cov(Y,X)$$

C. La variance d'une variable T est égale à la covariance de T avec elle-même : $var(T) = cov(T,T)$

$$cov(T,T) = \frac{\sum(t_i - \bar{t})\sum(t_i - \bar{t})}{n} = \frac{\sum(t_i - \bar{t})^2}{n} = Var(T)$$

I - Régression et corrélation

Exercices

D. La variance du triple d'une variable vaut le nonuple de la variance : $\text{var}(3T) = 9 \text{ var}(T)$

$$\text{Var}(3T) = \frac{\sum (3t_i - 3\bar{t})^2}{n} = \frac{\sum [3 * (t_i - \bar{t})]^2}{n} = 9 * \frac{\sum (t_i - \bar{t})^2}{n} = 9 \text{ Var}(T)$$

E. La covariance entre une variable T et son opposée -T est toujours négative

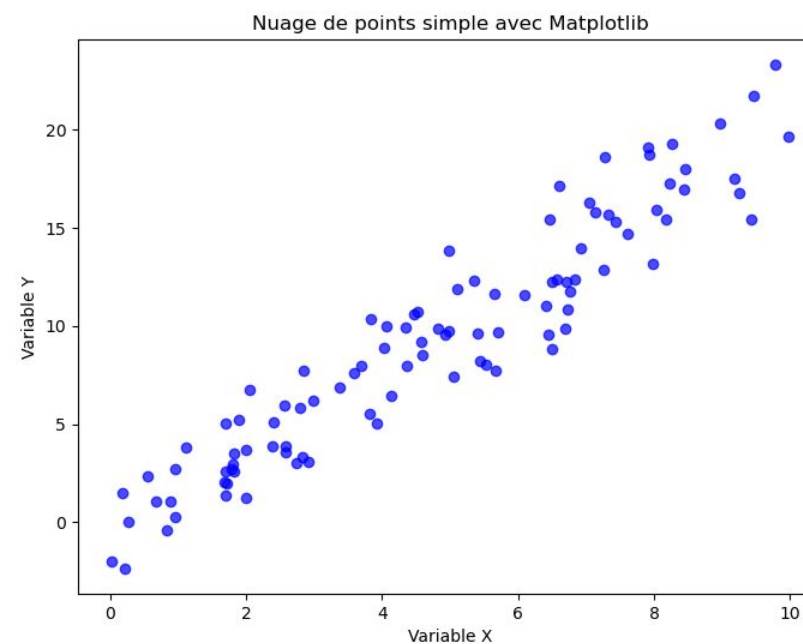
$$\text{Cov}(T, -T) = \frac{\sum (t_i - \bar{t}) \sum (-1 * (t_i - \bar{t}))}{n} = \frac{-1 * \sum (t_i - \bar{t}) \sum (t_i - \bar{t})}{n} = - \frac{\sum (t_i - \bar{t})^2}{n}$$

I - Régression et corrélation

Concept

Soit 2 variables X et Y **quantitatives + appariées + ayant ou non la même unité**. On cherche une relation entre X et Y.

Pouvons-nous tracer une droite ?
Est-ce qu'on peut écrire $f(X) = Y$?
Exprimer l'un à partir de l'autre ?



REGRESSION

$$\hat{Y} = f(X)$$

$$\hat{Y} = aX + b$$

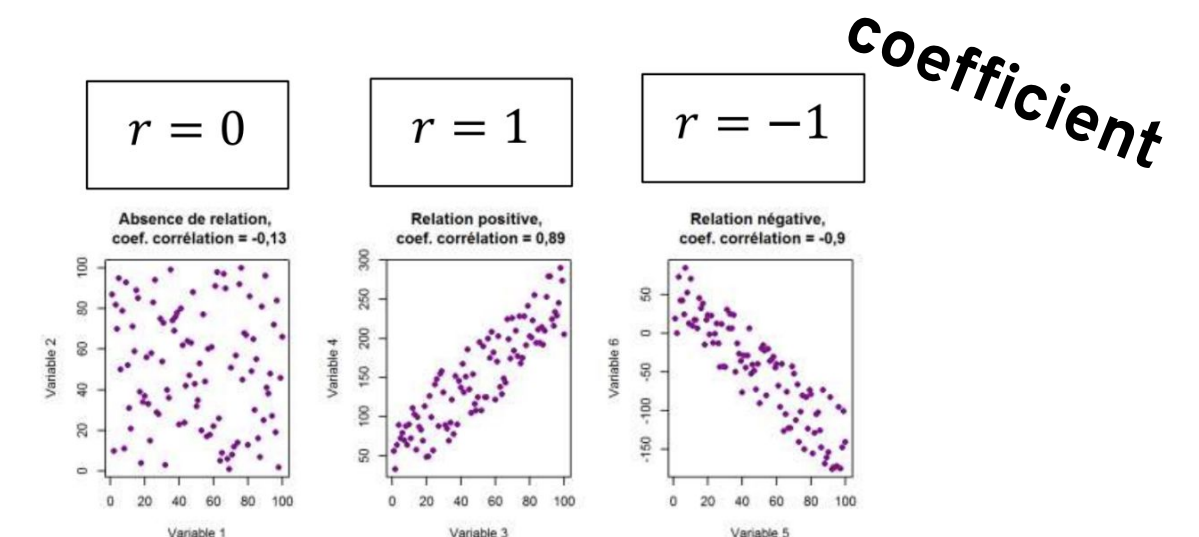
droite

Asymétrie : on ne peut pas permuter X et Y

$$f(X)=Y \neq f(Y)=X$$

A quel degré les 2 variables sont-elles liées (lien fort ou lien faible ???)

CORRELATION



Symétrique : X et Y peuvent être permuté

I - Régression et corrélation

Deux approches différentes

Régression	Correlation
○ Modélisation d'une droite ○ Apparition de la fonction $y = f(x)$ ○ X et Y ne sont pas interchangeables	○ Pas de calcul de droite ○ La valeur de r (coefficient de corrélation) est la même qu'on ait X en abscisse et Y en ordonnée ou l'inverse. ○ X et Y jouent exactement le même rôle

I - Régression et corrélation

Corrélation

Le coefficient de corrélation = coefficient de Pearson

$$r = \frac{cov(X, Y)}{\sqrt{Var(X) * Var(Y)}}$$

La covariance donne une indication sur la liaison entre X et Y

Le dénominateur permet de standardiser les valeurs : il transforme la covariance (*qui dépend des unités*) en une grandeur sans dimension comprise entre -1 et 1

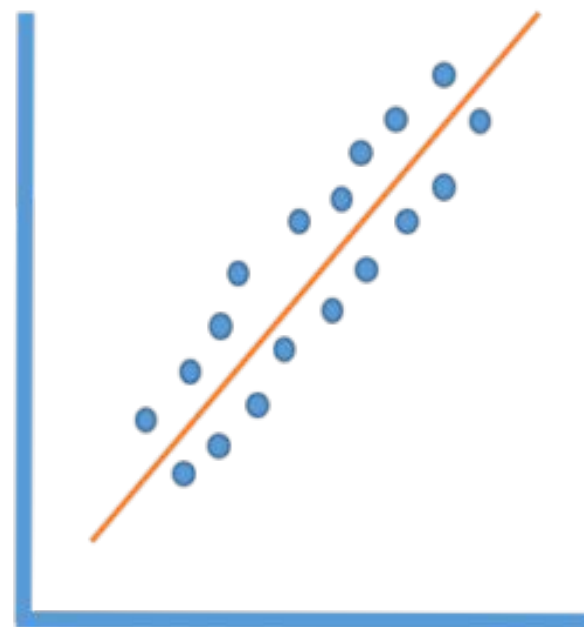
Le coefficient de corrélation r est un nombre compris entre -1 et 1 qui mesure la relation entre les deux variables x et y . Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation linéaire entre les variables est forte.

- **Si $r > 0$** , les valeurs prises par y ont tendance à croître quand les valeurs de x augmentent.
- **Si $r < 0$** , les valeurs prises par y ont tendance à décroître quand les valeurs de x augmentent.
- **Si $r = 0$** , les variations des variables x et y sont indépendantes.

I - Régression et corrélation

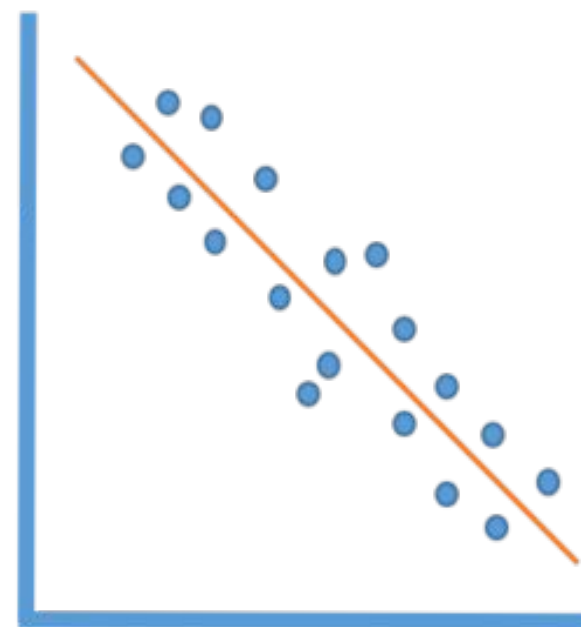
Corrélation

$$r > 0$$



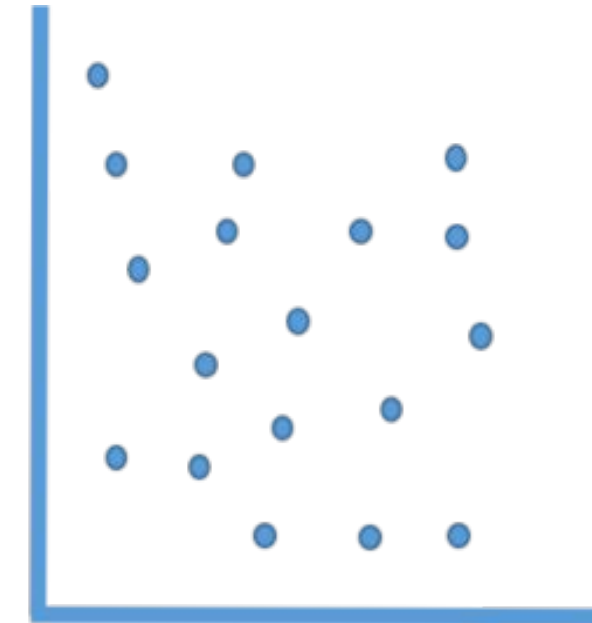
Positive Correlation

$$r < 0$$



Negative Correlation

$$r = 0$$



No Correlation

I - Régression et corrélation

Corrélation

Estimation dans un échantillon du coef.
de corrélation

$$r = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X) * \text{Var}(Y)}} \longrightarrow r = \frac{S_{XY}}{\sqrt{S_X^2 * S_Y^2}}$$

$$r = \frac{\frac{\sum(x_i - \bar{x}) \sum(y_i - \bar{y})}{n - 1}}{\sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1} * \frac{\sum(y_i - \bar{y})^2}{n - 1}}}$$

$$r = \frac{\sum(x_i - \bar{x}) \sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}}$$

$$r = \frac{\frac{\sum(x_i y_i) - n * \bar{x} \bar{y}}{n - 1}}{\sqrt{\frac{\sum(x_i^2) - n \bar{x}^2}{n - 1} * \frac{\sum(y_i^2) - n \bar{y}^2}{n - 1}}}$$

$$r = \frac{\sum(x_i y_i) - n * \bar{x} \bar{y}}{\sqrt{(\sum(x_i^2) - n \bar{x}^2) * (\sum(y_i^2) - n \bar{y}^2)}}$$

I - Régression et corrélation

Exercices

Une étude cherche à analyser l'existence d'un lien entre :

- X : le temps d'écran journalier (h/jour)
- Y : le degré de correction des lunettes (en dioptries, valeur absolue)

On recueille les données suivantes auprès de 10 adultes volontaires :

$$\bar{X} = 5,5 \quad ; \quad \bar{Y} = 1,5 \quad ; \quad \sum x_i^2 = 385 \quad ; \quad \sum y_i^2 = 27,24 \quad ; \quad \sum x_i y_i = 101,7$$

A- Le coefficient X en fonction de Y n'est pas le même que Y en fonction de X : $\rho(X,Y) \neq \rho(Y,X)$

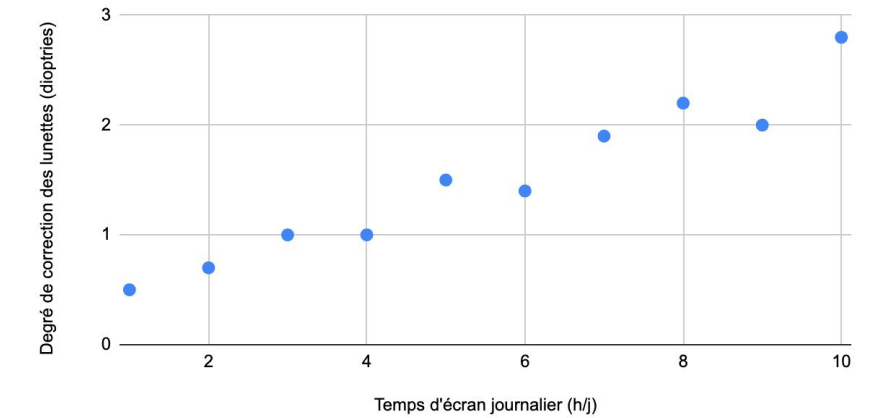
B- Le coefficient de Pearson a une valeur de 0,30

C- Le coefficient de Pearson a une valeur de 0,97

D- Le coefficient de Pearson a une valeur de 0,65

E- Il existe une liaison positive entre X et Y

Degré de correction des lunettes (dioptries) par rapport à
Temps d'écran journalier (h/j)



I - Régression et corrélation

10 Adultes volontaire $\rightarrow n=10$

Exercices

$$\bar{X} = 5,5 \ ; \ \bar{Y} = 1,5 \ ; \ \sum x_i^2 = 385 \ ; \ \sum y_i^2 = 27,24 \ ; \ \sum x_i y_i = 101,7$$

A- Le coefficient X en fonction de Y n'est pas le même que Y en fonction de X : $\rho(X,Y) \neq \rho(Y,X)$

Il s'agit d'un coefficient $\rightarrow$ il est identique et symétrique, on dit que $\rho(X,Y)=\rho(Y,X)$

B- Le coefficient de Pearson a une valeur de 0,30

C- Le coefficient de Pearson a une valeur de 0,97

D- Le coefficient de Pearson a une valeur de 0,65

On utilise la formule du cours :

$$r = \frac{\sum(x_i y_i) - n * \bar{x} \bar{y}}{\sqrt{(\sum(x_i^2) - n \bar{x}^2) * (\sum(y_i^2) - n \bar{y}^2)}} \longrightarrow r = \frac{101,7 - 10 * 5,5 * 1,5}{\sqrt{(385 - 10 * 5,5^2) * (27,24 - 10 * 1,5^2)}} \approx 0,97$$

$\downarrow$

E- Il existe une liaison positive entre X et Y

Attention à ne pas oublier de mettre la moyenne au carré!
(si tu oublies tu trouves 0,3)

I - Régression et corrélation

Corrélation

Test du coefficient de corrélation (*est-il représentatif ?*)

1] Hypothèses

- H0 : absence de liaison entre X et Y ($\rho_{X,Y}=0$)
- H1 : présence d'une liaison entre X et Y ($\rho_{X,Y}\neq 0$)

2] Conditions

- Indépendance des observations
- Liaison linéaire entre X et Y (*visualise-t-on une droite ?*)
- Bi-normalité de X et de Y

3] Calcul de la statistique

$$t_0 = \frac{r}{S_r}$$
$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$
$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$


4] Valeur Seuil

Table de Student à n-2 ddl

5] Conclusion

$|t_0| > t(n-2)_{ddl}$
REJET de H0

I - Régression et corrélation

Exercices

Une étude cherche à analyser l'existence d'un lien entre :

- X : le temps d'écran journalier (h/jour)
- Y : le degré de correction des lunettes (en dioptries, valeur absolue)

Précédemment nous avons calculé le coefficient de corrélation $r=0,97$, nous avons 10 participants et nous testons le coefficient avec un risque alpha de 5%.

A- L'hypothèse nulle H_0 s'écrit "la corrélation linéaire est nulle dans la population".

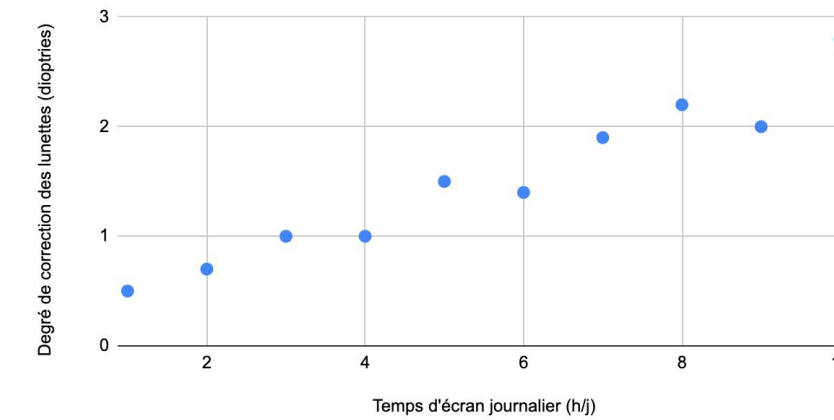
B- On effectue un test à 10ddl.

C- La valeur statistique seuil, lue dans la table, est 2,306.

D- La statistique vaut environ 11,3

E- On accepte H_0 , la corrélation est hautement significative.

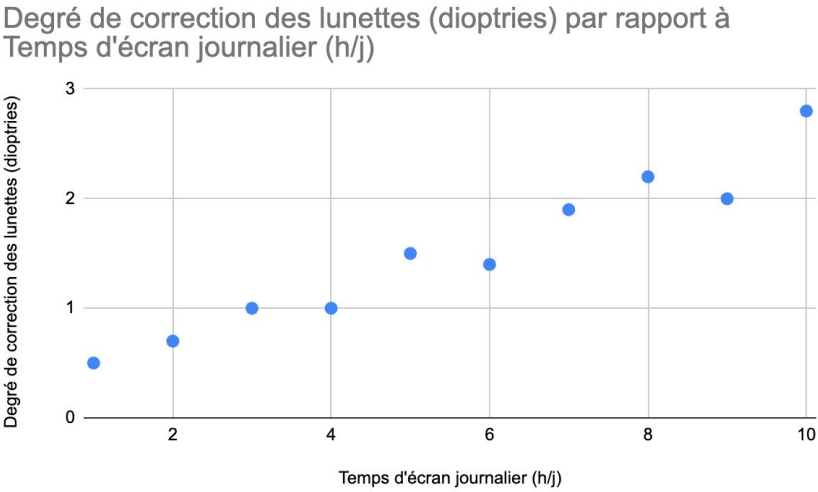
Degré de correction des lunettes (dioptries) par rapport à
Temps d'écran journalier (h/j)



I - Régression et corrélation

Exercices

r=0,97 n=10 participants alpha=5%



A- L'hypothèse nulle H0 s'écrit "la corrélation linéaire est nulle dans la population".

H0: $\rho_{X,Y}=0 \rightarrow$ la corrélation linéaire est nulle dans la population
H1: $\rho_{X,Y}\neq 0 \rightarrow$ la corrélation linéaire est significative dans la population

B- On effectue un test à 10ddl.

Sous H0 la statistique suit une loi de student à $n-2ddl = 10-2 = 8ddl$

C- La valeur statistique seuil, lue dans la table, est 2,306.

On lit dans la table de student, risque alpha = 5% et 8ddl

D- La statistique vaut environ 11,3

d.d.l. \ α	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228



$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,97 * \sqrt{10-2}}{\sqrt{1-0,97^2}} \approx 11,3$$

E- On accepte H0, la corrélation est hautement significative.

11,3 >> 2,306

→ On rejette H0, la corrélation est hautement signification

I - Régression et corrélation

$$y = f(x) = a + bx$$

Régression linéaire

- L'objectif est de prédire Y en fonction de X

Critère, Variable à expliquer

← $\hat{Y} = \alpha + \beta X + \varepsilon$ → Résidu = écart au modèle

Ordonnée à l'origine

pente de la droite

Prédicteur, variable explicative

La pente de la droite $\beta = \frac{cov(X, Y)}{Var(X)}$ *estimé par* $b = \frac{\sum(x_i - \bar{x}) \sum(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

L'ordonnée à l'origine $\alpha = Y - \beta X$ *estimé par* $a = \bar{y} - b\bar{x}$

I - Régression et corrélation

Régression linéaire

Test de la pente de la droite de régression (*est-elle représentative ?*)

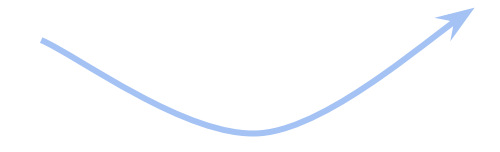
1] Hypothèses

- H0 : absence de liaison entre X et Y (pente $b=0$)
- H1 : Y dépend de X (pente $b \neq 0$)

2] Conditions

- Indépendance des observations
- Liaison linéaire entre X et Y (*visualise-t-on une droite ?*)
- Bi-normalité de X et de Y

3] Calcul de la statistique

$$t_0 = \frac{b}{S_b}$$
$$S_b = \sqrt{\frac{\frac{S_Y^2}{S_X^2} - b^2}{n - 2}}$$
$$t_0 = \frac{b}{\sqrt{\frac{\frac{S_Y^2}{S_X^2} - b^2}{n - 2}}}$$


4] Valeur Seuil

Table de Student à $n-2$ ddl

5] Conclusion

$|t_0| > t(n-2)_{ddl}$
REJET de H0

I - Régression et corrélation

Exercices

- A. Dans une approche de régression linéaire, le modèle permet de prédire une valeur unique du prédicteur pour chaque valeur du critère
- B. Si le coefficient de corrélation linéaire est positif, les pentes des droites de régression linéaire seront positives
- C. La pente dans la droite de régression linéaire s'exprime avec l'unité de critère
- D. Dans le modèle linéaire simple, les écarts au modèle correspondent aux résidus
- E. Plus le coefficient de corrélation est proche de 1, en valeur absolue, plus la liaison entre les variables est importante

I - Régression et corrélation

Exercices

- A. Dans une approche de régression linéaire, le modèle permet de prédire une valeur unique du prédicteur pour chaque valeur du critère**
→ Dans une approche de régression linéaire, le modèle permet de prédire une valeur unique du critère pour chaque valeur du prédicteur
- B. Si le coefficient de corrélation linéaire est positif, les pentes des droites de régression linéaire seront positives**
→ En effet, quand $r > 0$, y croît quand x croît, la pente est donc positive
- C. La pente dans la droite de régression linéaire s'exprime avec l'unité de critère**
→ La pente de la droite de régression correspond au coefficient α qui s'exprime dans l'unité de Y / l'unité X .

I - Régression et corrélation

Exercices

D. Dans le modèle linéaire simple, les écarts au modèle correspondent aux résidus.

E. Plus le coefficient de corrélation est proche de 1, en valeur absolue, plus la liaison entre les variables est importante.

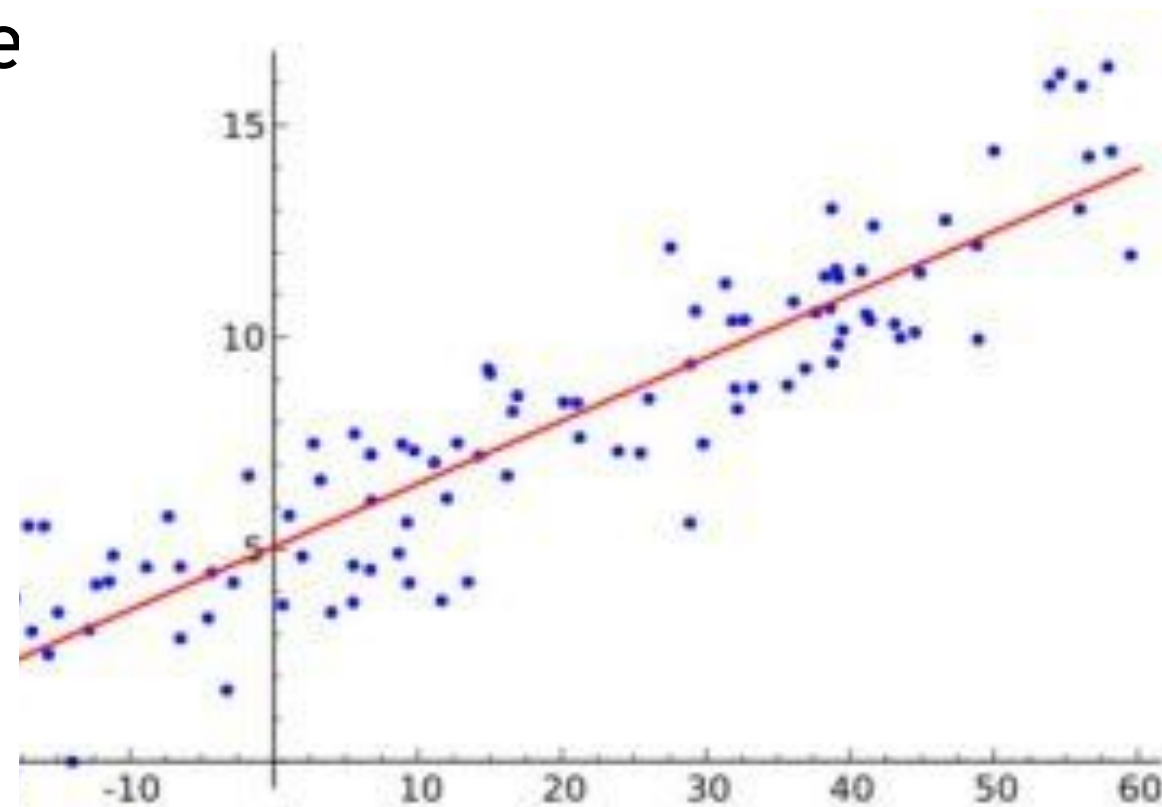
→ Effectivement plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation linéaire entre les variables est forte

I - Régression et corrélation

Exercices

Le dosage d'une substance sérique a été réalisée chez un échantillon représentatif de patients traités pour une pathologie P. Deux dosages ont été réalisés, à respectivement J0 (variable x en mmol/L) et J30 (variable y en mmol/L). Entre J0 et J30, les sujets ont été traités avec le traitement A.

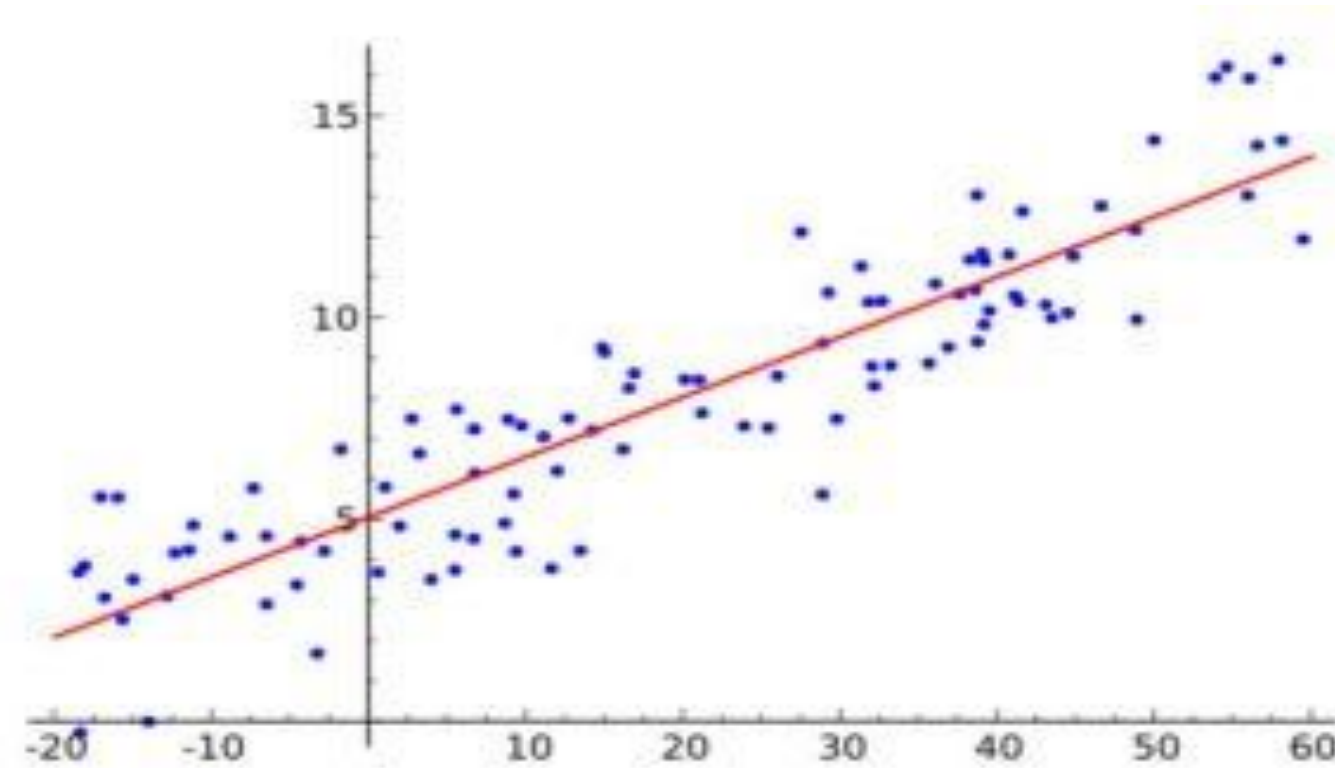
La figure jointe représente le nuage de points correspondant aux valeurs recueillies et à la droite de régression de y en fonction de



$$y=5+0,17x$$

I - Régression et corrélation

Exercices



$$y=5+0,17x$$

- A. Si on représentait la droite de régression de x en fonction de y , la pente serait de signe négatif
- B. On peut affirmer que la covariance $(x;y)$ est positive
- C. L'ordonnée à l'origine dans l'équation de la droite est compatible avec la figure
- D. Les valeurs des dosages à J0 et à J30 sont indépendantes
- E. Si un sujet présentait un dosage à J0 de 35 mmol/L, la valeur prédite de son dosage à J30 serait de 10 mmol/L

I - Régression et corrélation

Exercices

A. Si on représentait la droite de régression de x en fonction de y , la pente serait de signe négatif

→ D'après l'allure de la courbe, y croît quand x croît, autrement dit x croît quand y croît, la pente resterait de signe positif

B. On peut affirmer que la covariance $(x;y)$ est positive

→ y croît quand x croît, le coefficient de corrélation semble être proche de 1, il est donc bien positif donc la covariance est aussi positive

C. L'ordonnée à l'origine dans l'équation de la droite est compatible avec la figure

Rappel: L'ordonnée à l'origine se lit à l'intersection de la droite avec l'axe des ordonnées
soit : $\beta = 5$

I - Régression et corrélation

Exercices

D. Les valeurs des dosages à J0 et à J30 sont indépendantes

→ Au vu du coefficient de corrélation élevé, les deux dosages sont fortement corrélées donc dépendantes l'une de l'autre

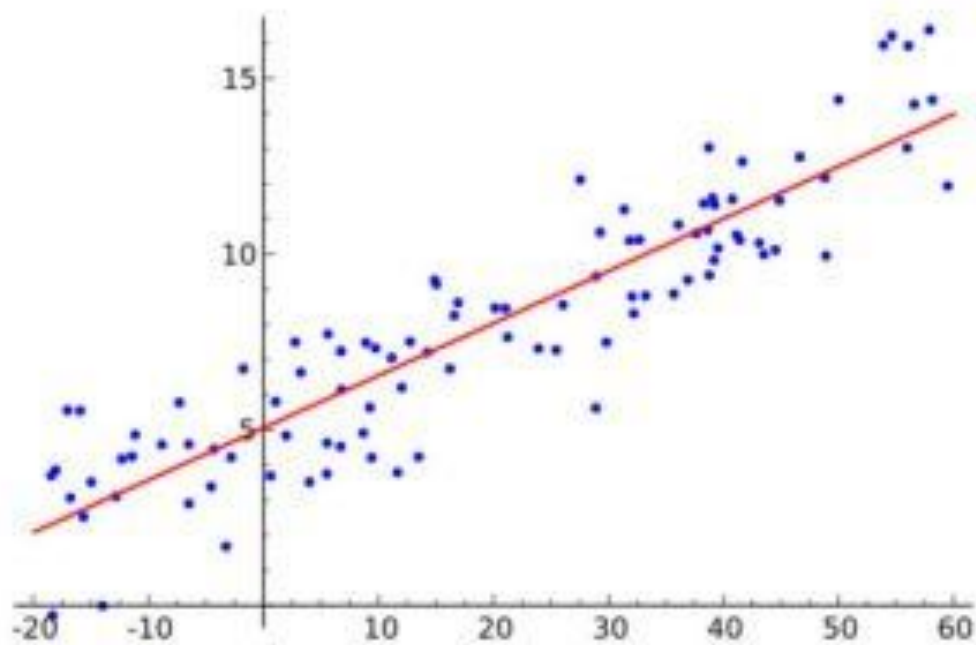
E. Si un sujet présentait un dosage à J0 de 35 mmol/L, la valeur prédite de son dosage à J30 serait de 10 mmol/L

→ Lecture graphique J0 en abscisse et J30 en ordonné OU par calcul :

$$Y = 5 + 0,17 \times 35 = 5 + 5,95 = 10,95$$

I - Régression et corrélation

Exercices



- A. Pour analyser la relation entre les dosages à J0 et J30, la démarche la plus adaptée est la régression linéaire
- B. Pour analyser la relation entre les dosages à J0 et J30, la démarche la plus adaptée est une étude de corrélation
- C. Pour analyser la relation entre les dosages à J0 et J30, la démarche la plus adaptée est la comparaison de moyenne sur série appariée
- D. Pour comparer les valeurs de dosages à J0 et J30, la démarche la plus adaptée est une étude de corrélation
- E. Pour comparer les valeurs de dosages à J0 et J30, la démarche la plus adaptée est la comparaison de moyenne sur série appariée

I - Régression et corrélation

Exercices

- A. Pour analyser la relation entre les dosages à J0 et J30, la démarche la plus adaptée est la régression linéaire
- B. Pour analyser la relation entre les dosages à J0 et J30, la démarche la plus adaptée est une étude de corrélation
→ Indice = “relation” donc test de **correlation**.
- C. Pour analyser la relation entre les dosages à J0 et J30, la démarche la plus adaptée est la comparaison de moyenne sur série appariée

I - Régression et corrélation

Exercices

D. Pour comparer les valeurs de dosages à J0 et J30, la démarche la plus adaptée est une étude de corrélation

E. Pour comparer les valeurs de dosages à J0 et J30, la démarche la plus adaptée est la comparaison de moyenne sur série appariée

→ On fait 2 mesures différentes sur un même échantillon, on est bien sur une série appariée. De plus on cherche à comparer les valeurs de dosages, une comparaison de moyenne est ici adaptée

I - Régression et corrélation

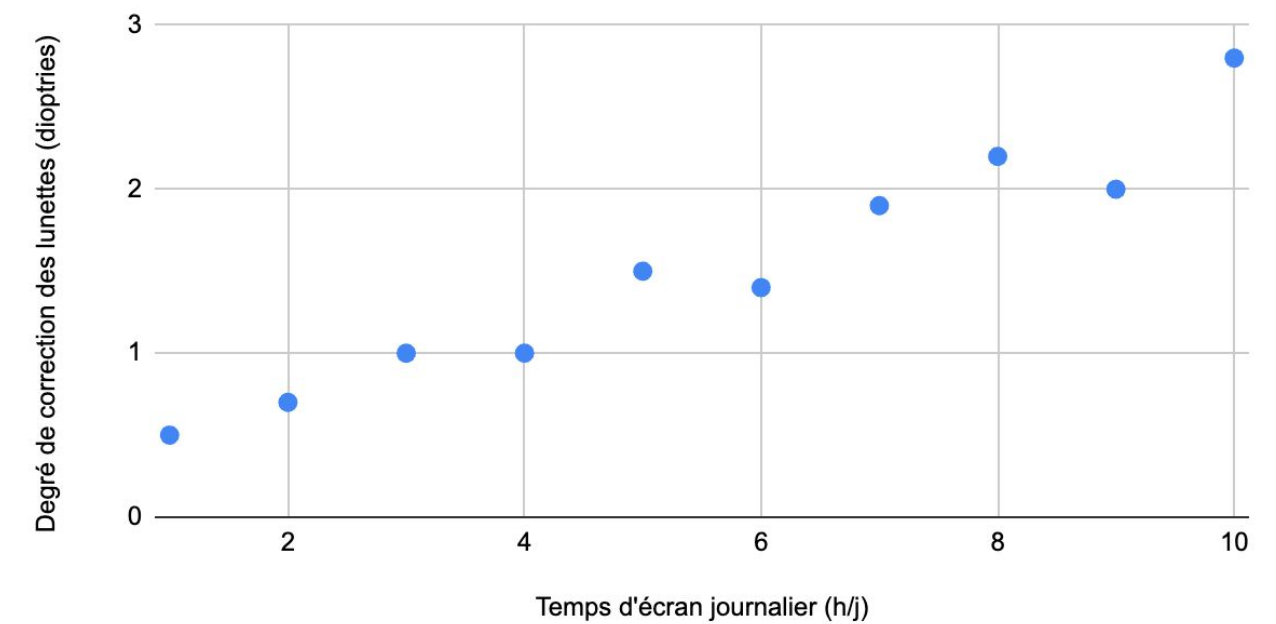
Exercices

Une étude cherche à analyser l'existence d'un lien entre :

- X : le temps d'écran journalier (h/jour)
- Y : le degré de correction des lunettes (en dioptries, valeur absolue)

On recueille les données suivantes auprès de 10 adultes volontaires :

Degré de correction des lunettes (dioptries) par rapport à Temps d'écran journalier (h/j)



$$\bar{X} = 5,5 \ ; \ \bar{Y} = 1,5 \ ; \ \sum x_i^2 = 385 \ ; \ \sum y_i^2 = 27,24 \ ; \ \sum x_i y_i = 101,7$$

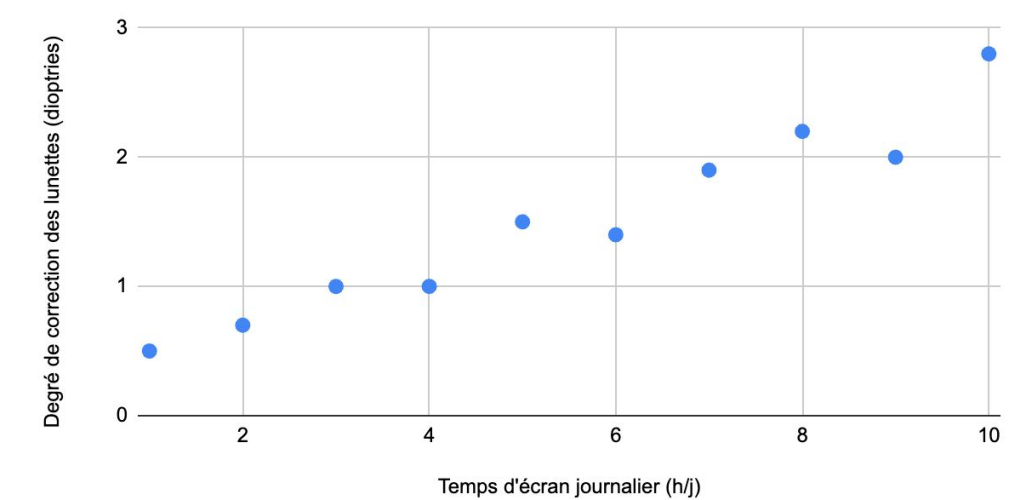
- A- La pente de la droite de Y en fonction de X est la même que la droite de X en fonction de Y
- B- La pente de la droite de Y en fonction de X vaut 0,23
- C- L'ordonnée à l'origine de Y en fonction de X vaut environ 0,235
- D- La droite prédit une correction de 3 dioptries si une personne est exposée à 10h d'écran journalier
- E- La droite prédit une correction de 2,5 dioptries si une personne est exposée à 10h d'écran journalier

I - Régression et corrélation

Exercices

$$\bar{X} = 5,5 \ ; \ \bar{Y} = 1,5 \ ; \ \sum x_i^2 = 385 \ ; \ \sum y_i^2 = 27,24 \ ; \ \sum x_i y_i = 101,7$$

Degré de correction des lunettes (dioptries) par rapport à Temps d'écran journalier (h/j)



A- La pente de la droite de Y en fonction de X est la même que la droite de X en fonction de Y

La régression linéaire est asymétrique, les paramètres de la droite X en fonction de Y ne sont pas les memes que ceux de la droite Y en fonction de X.

B- La pente de la droite de Y en fonction de X vaut 0,23

$$b = \frac{cov(X,Y)}{var(X)} = \frac{\sum(x_i y_i) - n * \bar{x} \bar{y}}{\sum(x_i^2) - n \bar{x}^2} = \frac{101,7 - 10 * 5,5 * 1,5}{385 - 10 * 5,5^2} = 0,23$$

C- L'ordonnée à l'origine de Y en fonction de X vaut environ 0,235

$$a = \bar{Y} - a\bar{X} = 1,5 - 0,23 * 5,5 = 0,235$$

$$Y = 0,235 + 0,23 * X$$

$$Y = 0,235 + 0,23 * 10 = 0,235 + 2,3 = 2,535$$

D- La droite prédit une correction de 3 dioptries si une personne est exposée à 10h d'écran journalier

E- La droite prédit une correction de 2,5 dioptries si une personne est exposée à 10h d'écran journalier

I - Régression et corrélation

Exercices

Lors de votre SIR (Stage Initiation à la Recherche), vous perdez la moitié de vos données... Il vous reste :

- L'équation de régression : $Y=2X-3$
- La variance de X : $\text{var}(X)=4$
- Le coefficient de corrélation : $r=0,8$
- La moyenne de Y : $m_y=7$

A- La moyenne de X vaut 5

B- La covariance(X,Y) vaut 16

C- Il existe une liaison positive entre X et Y : plus X augmente et plus Y diminue

D- La variance Y vaut 25

E- Si on exprime X en fonction de Y , alors la pente $b=8/25$

I - Régression et corrélation

Exercices

- L'équation de régression : $Y=2X-3$
- La variance de X : $\text{var}(X)=4$
- Le coefficient de corrélation : $r=0,8$
- La moyenne de Y : $m_y=7$

A- La moyenne de X vaut 5

B- La covariance(X,Y) vaut 16

La covariance(X,Y) vaut 8

C- Il existe une liaison positive entre X et Y : plus X augmente et plus Y diminue

Il existe une liaison positive entre X et Y (car $r=0,8$) : plus X augmente et plus Y **augmente**

D- La variance Y vaut 25

E- Si on exprime X en fonction de Y, alors la pente $b=8/25$

$$b = \text{cov}(X,Y) / \text{var}(Y) = 8/25$$

$$\hat{Y} = 2X - 3$$

$$b = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

$$2 = \frac{\text{cov}(X, Y)}{4}$$

$$\text{cov}(X, Y) = 8$$

$$a = \bar{Y} - b\bar{X}$$

$$-3 = 7 - 2\bar{X}$$

$$-10 = -2\bar{X}$$

$$\bar{X} = 5$$

$$r = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$$

$$0,8 = \frac{8}{\sqrt{4}\sqrt{\text{var}(Y)}}$$

$$0,8 = 4 * \frac{1}{\sqrt{\text{var}(Y)}}$$

$$\sqrt{\text{var}(Y)} = 5$$

$$\text{var}(Y) = 25$$

II - Comparaison proportions et moyenne

Exercices



Dans le laboratoire de Gru (Moi, moche et méchant), on teste deux nouvelles machines à confiture de banane. On veut savoir si la proportion de Minions satisfaits diffère entre les deux modèles.

Machine A : 200 minions testeurs ; 130 satisfaits

Machine B : 180 minions testeurs ; 135 satisfaits

- A- On compare des moyennes sur échantillons appariés.
- B- 75% des minions sont satisfaits par la machine B.
- C- Sous H_0 , la proportion combinée $p_0=0,65$.
- D- La statistique Z vaut environ -2,12.
- E- Avec un risque de 5%, on accepte H_0 : il n'y a pas de différence significative entre ces 2 machines.

II - Comparaison proportions et moyenne

Exercices



Machine A : 200 minions testeurs ; 130 satisfaits
Machine B : 180 minions testeurs ; 135 satisfaits

$$\longrightarrow p_A = \frac{130}{200} = 65\% \quad p_B = \frac{135}{180} = 75\%$$

A- On compare des moyennes sur échantillons appariés.

On compare des proportions sur échantillons indépendants.

B- 75% des minions sont satisfaits par la machine B.

C- Sous H_0 , la proportion combiné $p_0=0,65$.

$$p_0 = \frac{n_A p_A + n_B p_B}{n_A + n_B} = \frac{130 + 135}{200 + 180} = \frac{265}{380} \approx 0,697 \approx 70\%$$

$$Z = \frac{p_A - p_B}{\sqrt{p_0(1 - p_0) * \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{0,65 - 0,75}{\sqrt{0,7 * 0,3 * \left(\frac{1}{200} + \frac{1}{180}\right)}} = -2,12$$

D- La statistique Z vaut environ -2,12.

E- Avec un risque de 5%, on accepte H_0 : il n'y a pas de différence significative entre ces 2 machines.

On lit la valeur seuil dans la table de l'écart réduit avec alpha 0,5%, ce qui donne $Z(\text{table})=1,96$

$$|-2,12| \gg 1,96$$

$$Z_{\text{obs}} \gg Z(\text{table})$$

On rejette H_0 , la satisfaction diffère entre les 2 machines

II - Pourcentage

Comparaison de pourcentages [RAPPEL GDT 1]

Comparaison de 2 proportions indépendantes

$$H_0: \pi_1 = \pi_2 = \pi_0$$

Test du Khi 2

	Succès	Echec	
Groupe 1	O_{11}	O_{12}	n_1
Groupe 2	O_{21}	O_{22}	n_2
	n'_1	n'_2	N

Effectif Théorique $T_{ij} = \frac{n_i * n'_j}{N}$

Statistique
$$\chi^2_{obs} = \frac{(O_{11} - T_{11})^2}{T_{11}} + \frac{(O_{12} - T_{12})^2}{T_{12}} + \frac{(O_{21} - T_{21})^2}{T_{21}} + \frac{(O_{22} - T_{22})^2}{T_{22}}$$

Valeur Seuil

$$\chi^2_{th}(\alpha ; 1ddl) \text{ Table de Pearson}$$

Test Z

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n_1} + \frac{p_0 \cdot (1 - p_0)}{n_2}}} \text{ avec } p_0 = \frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2}$$

Valeur Seuil

$$Z_{th} = U_\alpha$$

table écart réduit

Conclusion

$$|z| < Z(\text{table}) \text{ ou } \chi < \chi(\text{table}) \rightarrow H_0 \text{ Acceptable}$$

II - Comparaison proportions et moyenne

Exercices



Au siège de Dunder Mifflin à Scranton, Michael Scott affirme que 70% des employés respectent les délais sans relance”. Pour vérifier cette affirmation, Jim Halpert réalise un audit interne. Sur un échantillon de 35 employés, 21 ont respecté les délais sans relance.

Est-ce que la réalité est cohérente avec les statistiques de Michael Scott, avec un seuil $\alpha = 5\%$?

- A- On peut réaliser un test du khi 2 d’homogénéité.
- B- La statistique Z vaut -1,29
- C- La statistique Z vaut -3,98
- D- On accepte H_0 : étonnamment l’audit interne de Jim ne révèle pas de différence significative avec la théorie
- E- La statistique du khi2 équivalente vaut 1,66

II - Comparaison proportions et moyenne

Exercices



Au siège de Dunder Mifflin à Scranton, Michael Scott affirme que 70% des employés respectent les délais sans relance”. Pour vérifier cette affirmation, Jim Halpert réalise un audit interne. Sur un échantillon de 35 employés, 21 ont respecté les délais sans relance.

Est-ce que la réalité est cohérente avec les statistiques de Michael Scott, avec un seuil $\alpha = 5\%$?

$$p_1 = \frac{21}{35} = 60\%$$

A- On peut réaliser un test du khi 2 d'homogénéité.

On peut réaliser un test du khi 2 d'ajustement/conformité

B- La statistique Z vaut -1,29

C- La statistique Z vaut -3,98

$$Z = \frac{p_1 - \pi_{th}}{\sqrt{\frac{\pi_{th}(1 - \pi_{th})}{n}}} = \frac{0,6 - 0,7}{\sqrt{\frac{0,7 * 0,3}{35}}} = -1,29$$

D- On accepte H0 : étonnamment l'audit interne de Jim ne révèle pas de différence significative avec la théorie

La valeur théorique lu dans la table de l'écart réduit est : $Z(5\%) = 1,95$

$$|-1,29| < 1,95 \rightarrow H_0 \text{ est acceptable}$$

E- La statistique du khi2 équivalente vaut 1,66 $Z^2 = 1,29^2 = 1,66 = \text{khi2}$

II - Pourcentage

Comparaison de pourcentages [RAPPEL GDT 1]

Comparaison à un pourcentage théorique :
 $H_0: \pi = \pi_{th}$

Test du Khi 2
(de conformité)

Observé	$O_1 = np$	$O_2 = n(1 - p)$
Théorique	$T_1 = n\pi_0$	$T_2 = n(1 - \pi_0)$

$$\chi^2_{obs} = \frac{(O_1 - T_1)^2}{T_1} + \frac{(O_2 - T_2)^2}{T_2}$$

Valeur Seuil

$\chi^2_{th}(\alpha ; 1ddl)$ Table de Pearson

Test Z

$$z = \frac{p - \pi_{th}}{\sqrt{\frac{\pi_{th} \cdot (1 - \pi_{th})}{n}}}$$

Valeur Seuil

$$Z_{th} = U_{\alpha}$$

table écart réduit

Conclusion
 $Z < Z(\text{table})$ ou $\chi < \chi(\text{table})$
→ H_0 Acceptable

II - Comparaison proportions et moyenne

Exercices



Dans le monde de Emmet Brickowski (personnage légo), un programme d’entraînement intensif a été mis au point pour améliorer la vitesse de construction des Maîtres Bâtisseurs. On mesure le temps (en minutes) nécessaire pour construire un vaisseau spatial avant et après l’entraînement chez 5 participants. Avec un risque de 5%, on cherche à savoir si l'entraînement est réellement efficace. La variance estimée vaut 0,8. Les valeurs suivent une loi normale.

Avant	18	16	20	15	19
Après	15	14	17	14	16

- A- On peut réaliser un test du khi 2 de McNemar.
- B- Sous H_0 , la différence de moyenne est nulle.
- C- La différence de moyenne calculé vaut 15,6.
- D- La statistique t observé vaut 6
- E- On conclut que l’entrainement est efficace pour réduire le temps de construction

II - Comparaison proportions et moyenne

Exercices



n = 5 participants. Avec un risque de 5% ; La variance estimée vaut 0,8.

Avant	18	16	20	15	19
Après	15	14	17	14	16

$$m_{AVANT} = \frac{18 + 16 + 20 + 15 + 19}{5} = \frac{88}{5} = 17,6$$

$$m_{APRES} = \frac{15 + 14 + 17 + 14 + 16}{5} = \frac{76}{5} = 15,2$$

A- On peut réaliser un test du khi 2 de McNemar.

On réalise un test de Student sur échantillon appariés. Le khi 2 de McNemar est utilisé pour comparer des *PROPORTIONS* appariées.

$$m_D = AVANT - APRES = 17,6 - 15,2 = 2,4$$

B- Sous H0, la différence de moyenne est nulle.

$$H_0: \mu_D = 0$$

C- La différence de moyenne calculé vaut 15,6.

Elle vaut 2,4

$$t_{obs} = \frac{m_D}{\sqrt{\frac{s^2}{n}}} = \frac{2,4}{\sqrt{\frac{0,8}{5}}} = 6$$

D- La statistique t observé vaut 6

E- On conclut que l'entraînement est efficace pour réduire le temps de construction

On lit dans la table du student 5% et n-1 ddl = 4ddl → 2,78

6 >> 2,78 → $t_{obs} > t(table)$ → REJET H_0 : il y a une vraie différence

III - Moyenne et variance

Test paramétrique [RAPPEL GDT 1]

Comparaison moyennes appariées
 $H_0: \mu = \mu_0 = 0$

On connaît la Variance de
la population σ^2

Test Z (de l'écart-réduit)

$$Z_{obs} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

Valeur Seuil

$$Z_{th} = U_\alpha \text{ de l'écart réduit}$$

$Z_{obs} < Z_{th}$ ou $t_{obs} < t_{th}$
H0 acceptable

On ne connaît pas σ^2

X suit une loi normale
OU bien $N > 30$

Test t de Student

$$t_{obs} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

Valeur Seuil

$$t_{th} = Student(\alpha; N - 1ddl)$$

I - Comparaison proportions et moyenne

Exercices



Bubba, l’ami de Forest Gump, vient d’ouvrir un restaurant spécialisé sur les crevettes. Il veut savoir si la taille moyenne des crevettes diffère entre deux lots provenant de fournisseurs différents (risque 5% Fischer et Student) :

	Lot A	Lot B
moyenne (cm)	5,3	5,7
effectif	25	50
variance estimée	1,2	1,1

- A- La statistique pris par le test de Fisher vaut 1,09
- B- La statistique lue dans la table de Fischer vaut environ 0,99
- C- La statistique lue dans la table de Fischer vaut environ 2,01
- D- On rejette l’Homoscédasticité.

I - Comparaison proportions et moyenne

Exercices



Bubba, l'ami de Forest Gump, vient d'ouvrir un restaurant spécialisé sur les crevettes. Il veut savoir si la taille moyenne des crevettes diffère entre deux lots provenant de fournisseurs différents (risque 5% Fischer et Student) :

	Lot A	Lot B
moyenne (cm)	5,3	5,7
effectif	25	50
variance estimée	1,2	1,1

A- La statistique pris par le test de Fisher vaut 1,09

On fait le rapport de la grande variance sur la petite variance $F_{obs} = \frac{s_A^2}{s_B^2} = \frac{1,2}{1,1} = 1,09$

B- La statistique lue dans la table de Fischer vaut environ 0,99

C- La statistique lue dans la table de Fischer vaut environ 2,01

$$F \begin{cases} \alpha/2 \\ v_1 = n_1 - 1ddl \\ v_2 = n_2 - 1ddl \end{cases} \rightarrow F \begin{cases} 0,05/2 \\ v_1 = 25 - 1ddl \\ v_2 = 50 - 1ddl \end{cases} \rightarrow F \begin{cases} 0,025 \\ v_1 = 24 \\ v_2 = 49 \end{cases} \approx 2,01$$

D- On rejette l'Homoscédasticité

Faux, on accepte H0 car Fobs < Fth

III - Moyenne et variance

Test de Fisher [RAPPEL GDT1]

Il permet de comparer la variance de 2 échantillons.

$H_0 : \sigma_1^2 = \sigma_2^2 \rightarrow$ les 2 échantillons sont issus de la même population (Homoscédasticité)

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Soit s_1^2 et s_2^2 des estimateurs non biaisés de la variance de nos échantillons $\rightarrow s_1^2 = \frac{n}{n-1} \text{Variance}$

$\Rightarrow$ la statistique est :

$$F_{obs} = \frac{s_1^2}{s_2^2}$$

Avec $s_1^2 > s_2^2$

$\Rightarrow$ la valeur seuil est :

$$F \begin{cases} \alpha/2 \\ v_1 = n_1 - 1ddl \\ v_2 = n_2 - 1ddl \end{cases}$$

$\Rightarrow$ la conclusion est :

$$F_{obs} < F$$

H_0 acceptable

I - Comparaison proportions et moyenne

Exercices



Bubba, l’ami de Forest Gump, vient d’ouvrir un restaurant spécialisé sur les crevettes. Il veut savoir si la taille moyenne des crevettes diffère entre deux lots provenant de fournisseurs différents (risque 5% Fischer et Student) :

	Lot A	Lot B
moyenne (cm)	5,3	5,7
effectif	25	50
variance estimée	1,2	1,1

- A- La variance commune
- B- La statistique Tobs
- C- Le degré de liberté
- D- Conclusion

I - Comparaison proportions et moyenne

Exercices



Bubba, l'ami de Forest Gump, vient d'ouvrir un restaurant spécialisé sur les crevettes. Il veut savoir si la taille moyenne des crevettes diffère entre deux lots provenant de fournisseurs différents (risque 5% Fischer et Student) :

	Lot A	Lot B
moyenne (cm)	5,3	5,7
effectif	25	50
variance estimée	1,2	1,1

A- La variance commune = 1,13

$$s_C^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{(25 - 1)1,2 + (50 - 1)1,1}{25 + 50 - 2} = 1,13$$

B- La statistique Tobs = -1,53

$$t_{obs} = \frac{m_A - m_B}{\sqrt{s_C^2 * \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{5,3 - 5,7}{\sqrt{1,13 * \left(\frac{1}{25} + \frac{1}{50}\right)}} = -1,53$$

C- Le degré de liberté = 75-2= 73ddl

D- Conclusion

III - Moyenne et variance

Test paramétrique [RECAP]

Comparaison moyennes indépendantes

$$H_0: \mu_1 = \mu_2 = \mu_0$$

Util si n_1 ou $n_2 < 30$

$$n_1 \geq 30 \\ \text{et } n_2 \geq 30$$

Test Z (de l'écart-réduit)

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Valeur Seuil

$$Z_{th} = U_\alpha \text{ de l'écart réduit}$$

X suit une loi normale
ET il y a une égalité des
variances (Homoscédasticité,
test Fischer)

Variance commune

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Test t de Student

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_c^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Valeur Seuil

$$t_{th} = Student(\alpha; n_1 + n_2 - 2ddl)$$

$$Z_{obs} < Z_{th} \text{ ou } t_{obs} < t_{th}$$

H0 acceptable

II - Tests diagnostiques

Caractéristiques Intrinsèques

Sensibilité	Spécificité
○ Probabilité de tests positifs chez les malades ○ Capacité d'un test à découvrir les malades	○ Probabilité de tests négatifs chez les non malades ○ Capacité d'un test à découvrir les sujets sains
$Se = P(T+/M+)$	$Sp = P(T-/M-)$

II - Tests diagnostiques

Caractéristiques extrinsèques

Valeur prédictive positive	Valeur prédictive négative
○ Probabilité d'être malade sachant le test est positif ○ Capacité d'un test à prédire la maladie	○ Probabilité de ne pas être malade sachant que le test est négatif ○ Capacité d'un test à prédire l'absence de maladie
$VPP = P(M+/T+)$	$VPN = P(M-/T-)$

→ Les VPP et VPN varient en fonction de la prévalence de la maladie dans la région. En effet, si p augmente, la VPP augmente et la VPN diminue. Un test n'aura pas la même valeur d'une région à une autre.

II - Tests diagnostiques

Indice de Youden

Fournit une indication de performance globale du test :

$$\text{Sensibilité} + \text{Spécificité} - 1$$

Il varie de -1 à 1, un test parfait aurait un indice de Youden à 1

II - Tests diagnostiques

Rapport de vraisemblance

Rapport de vraisemblance positif	Rapport de vraisemblance négatif
$L = Se/1-Sp$	$\lambda = 1-Se/Sp$

II - Tests diagnostiques

Estimation des paramètres

	Malade (M+)	Malade (M-)	Total
Test positif (T+)	VP	FP	VP+FP
Test négatif (T-)	FN	VN	FN+VN
Total	VP+FN	FP+VN	VP+FP+VN+FN

- Calcul de la prévalence de la maladie:

$$p = \frac{VP+FN}{N}$$

II - Tests diagnostiques

Estimation des paramètres

- Sensibilité : $Se = \frac{VP}{VP+FN}$

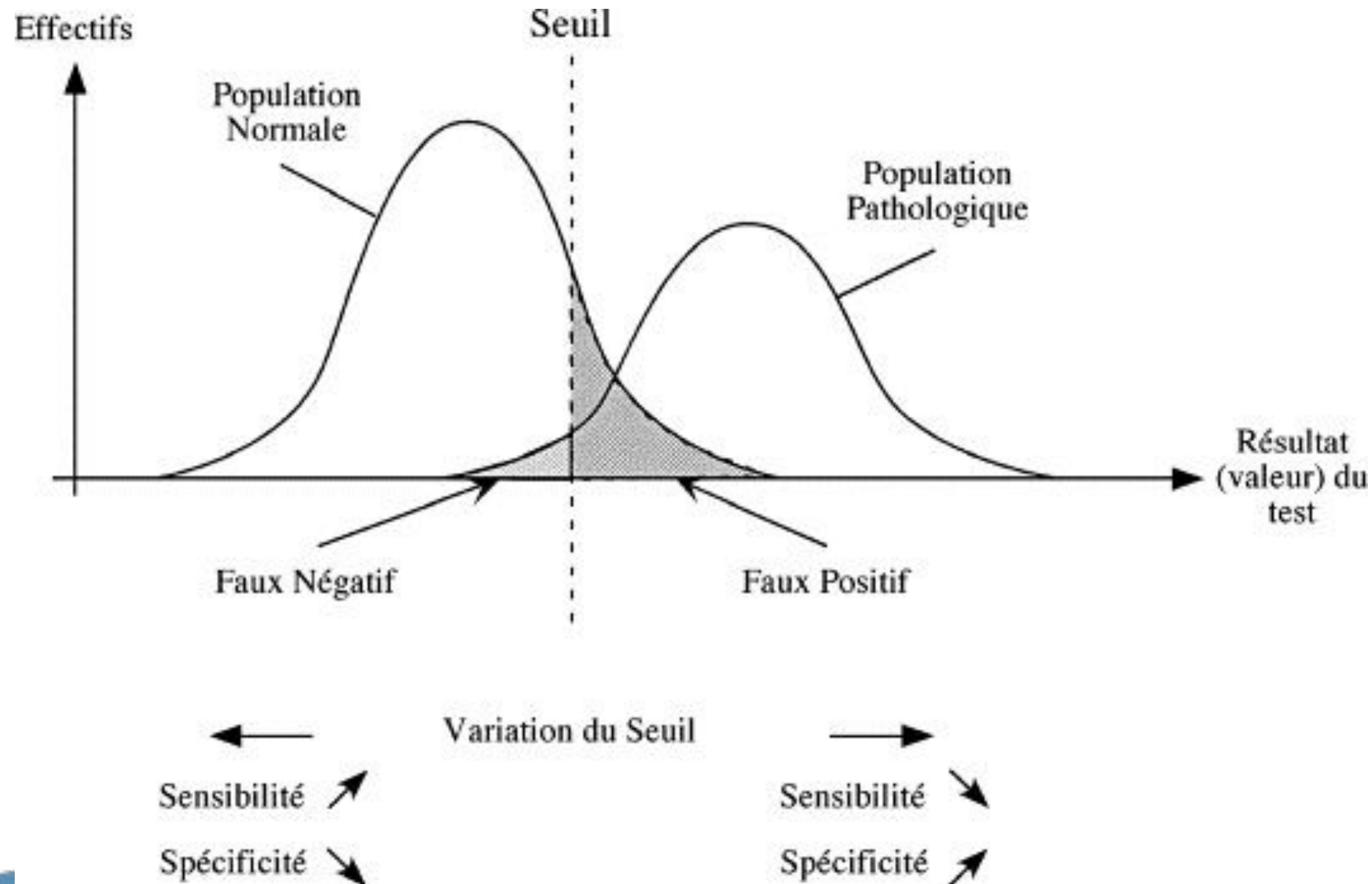
- Spécificité : $Sp = \frac{VN}{VN+FP}$

- Valeur prédictive positive : $VPP = \frac{VP}{VP+FP}$

- Valeur prédictive négative : $VPN = \frac{VN}{VN+FN}$

II - Tests diagnostiques

Variation de Sp et Se avec le seuil



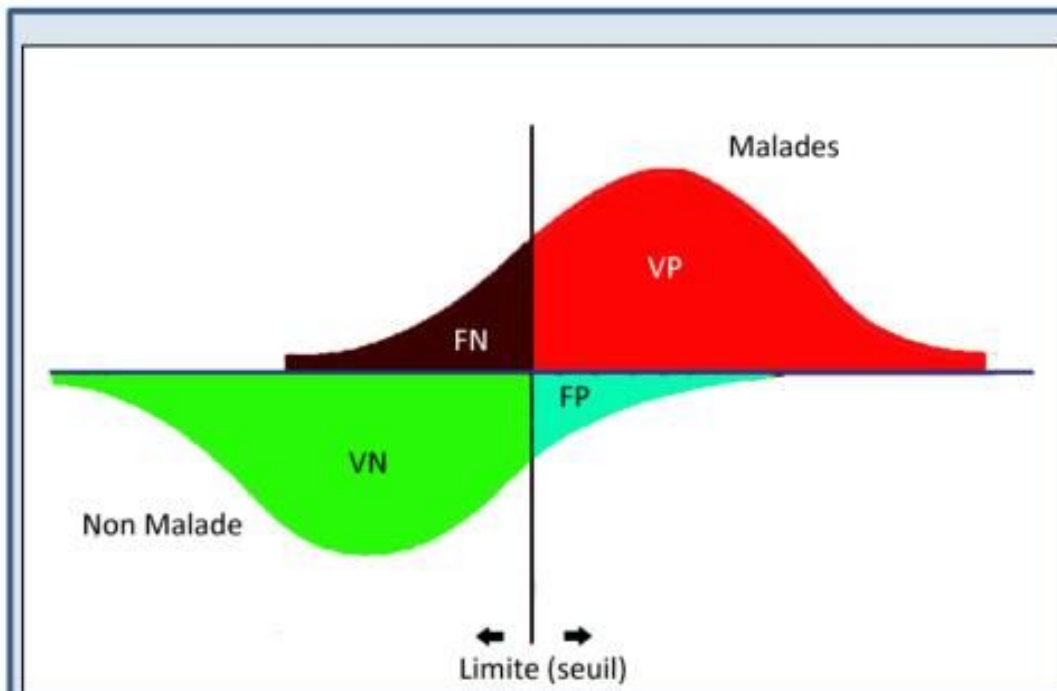
II - Tests diagnostiques

Impact sur la sensibilité	○ Si une valeur supérieure au seuil constitue un test positif : ➤ Lorsque le seuil augmente la sensibilité diminue ○ Si une valeur inférieure au seuil constitue un test positif : ➤ Lorsque le seuil augmente la sensibilité augmente
Impact sur la spécificité	○ Si une valeur supérieure au seuil constitue un test positif : ➤ Lorsque le seuil augmente la spécificité augmente ○ Si une valeur inférieure au seuil constitue un test positif : ➤ Lorsque le seuil augmente la spécificité diminue

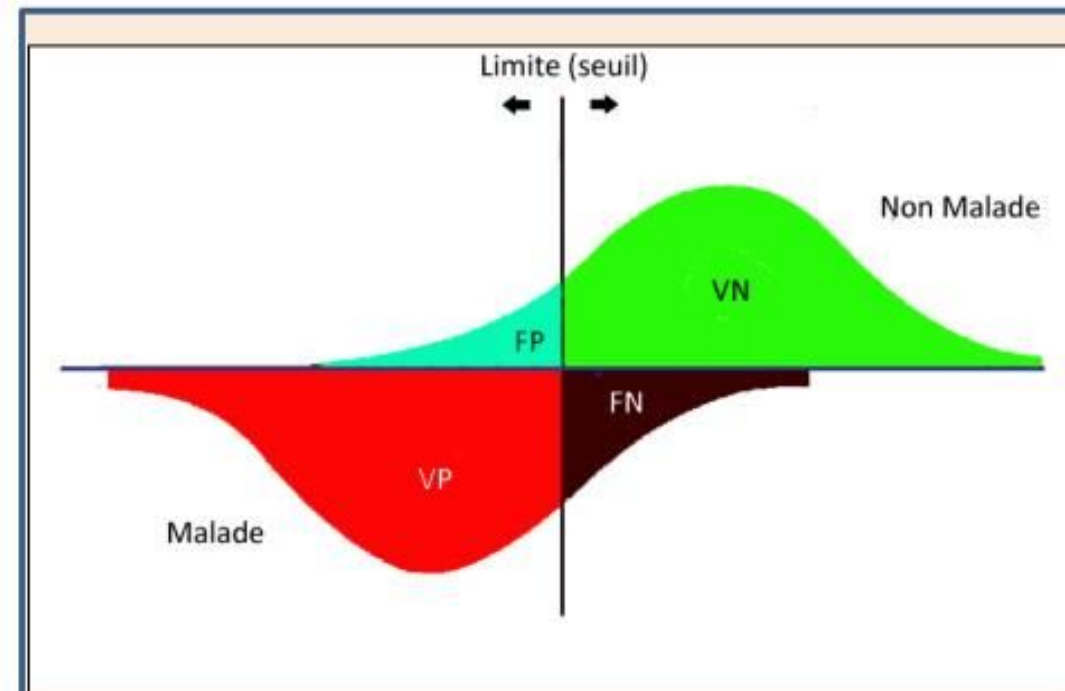
II - Tests diagnostiques



Variations du seuil



- Chez les malades, les valeurs sont plus élevées que chez les non malades (Diabète)
- Une augmentation de la valeur de la limite va :
 - Augmenter la spécificité
 - Diminuer la sensibilité



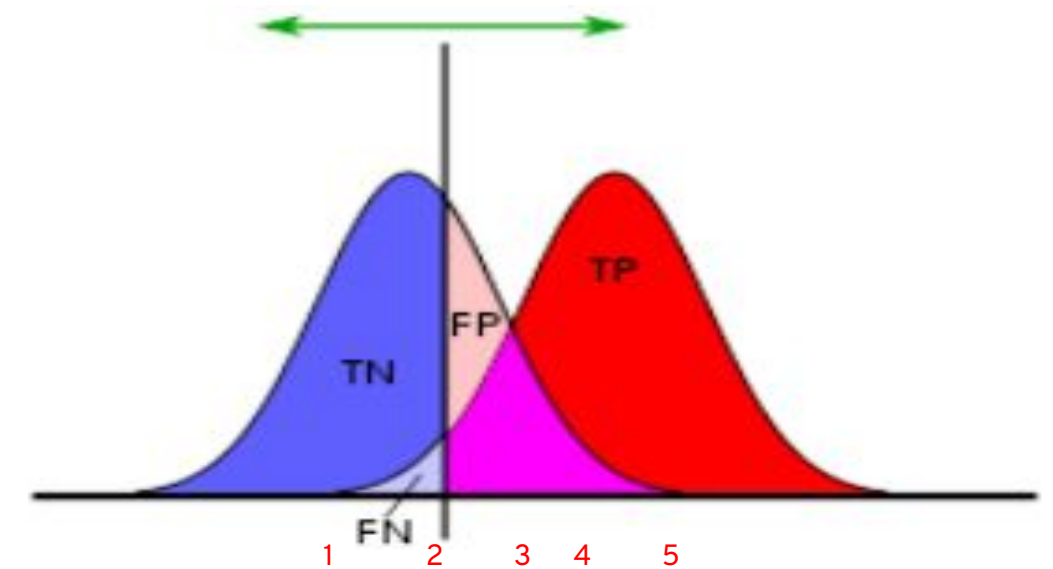
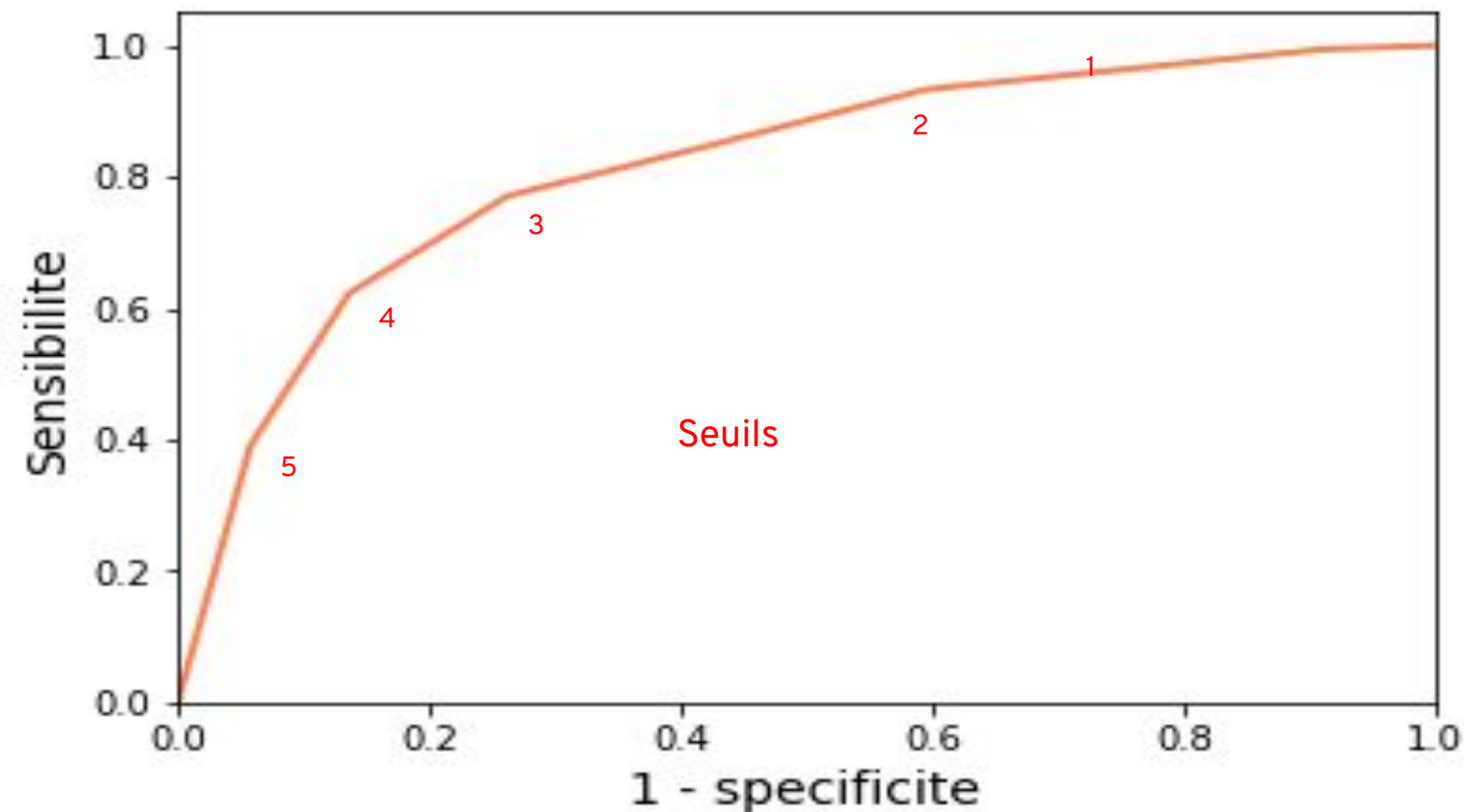
- Chez les malades, les valeurs sont moins élevées que chez les non malades (Hypothyroïdie)
- Une augmentation de la valeur de la limite va :
 - Diminuer la spécificité
 - Augmenter la sensibilité

III - Courbe ROC

ROC : Receiver Operating Characteristic

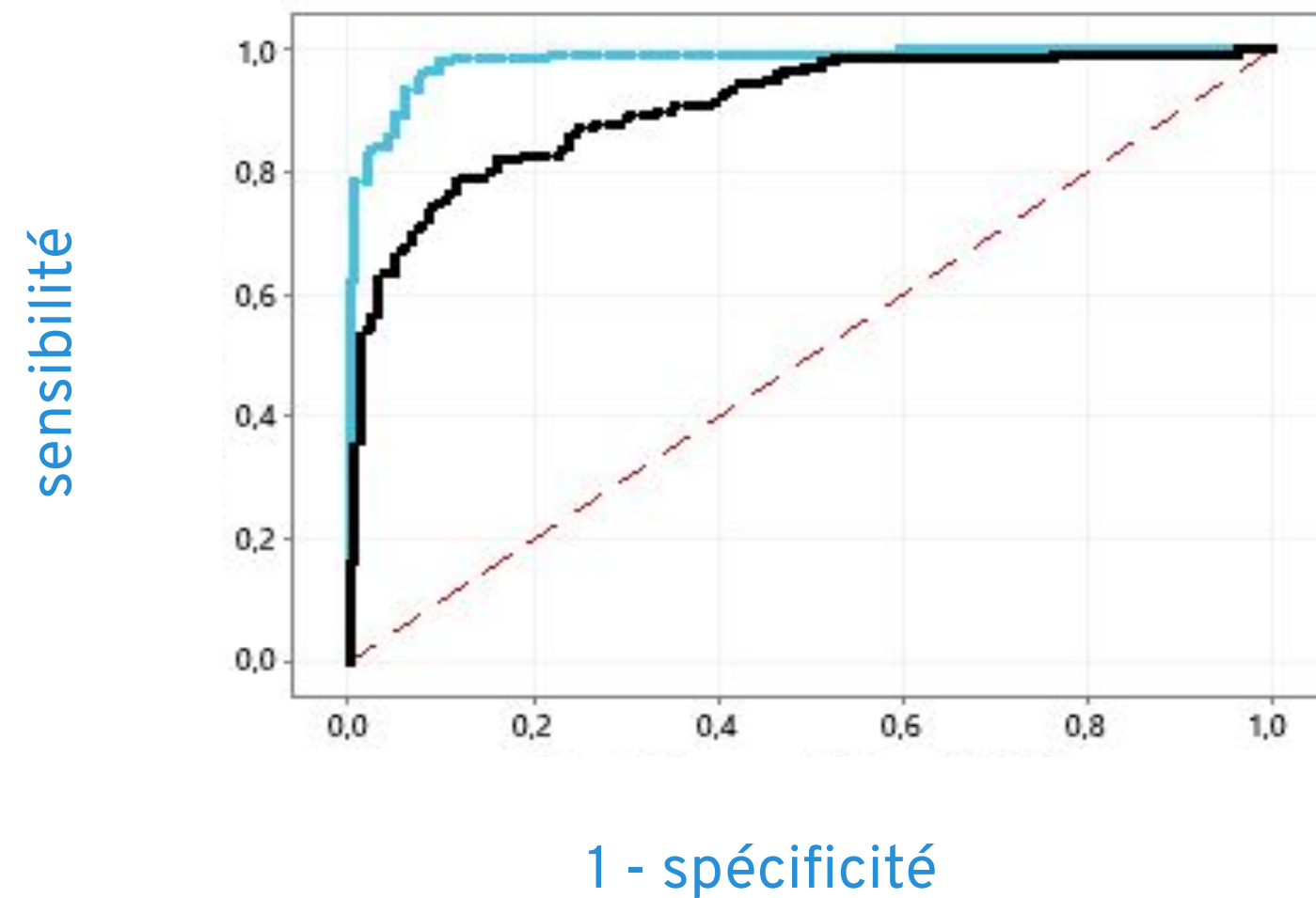
Courbe reliant la spécificité et la sensibilité d'un test diagnostique selon le choix de son seuil.

Exemple :



III - Courbe ROC

Plus l'aire sous la courbe du test est élevée, meilleur est le test diagnostique en question :



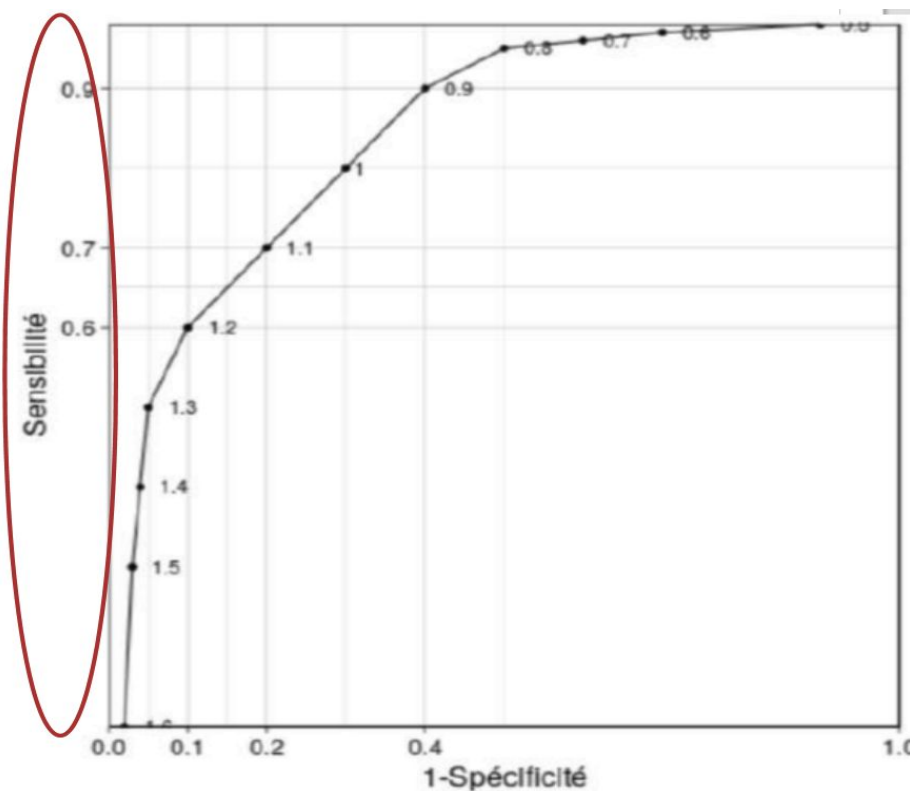
Ici, le test représenté par la courbe **bleue** est meilleur que celui de la courbe noire

III - Courbe ROC

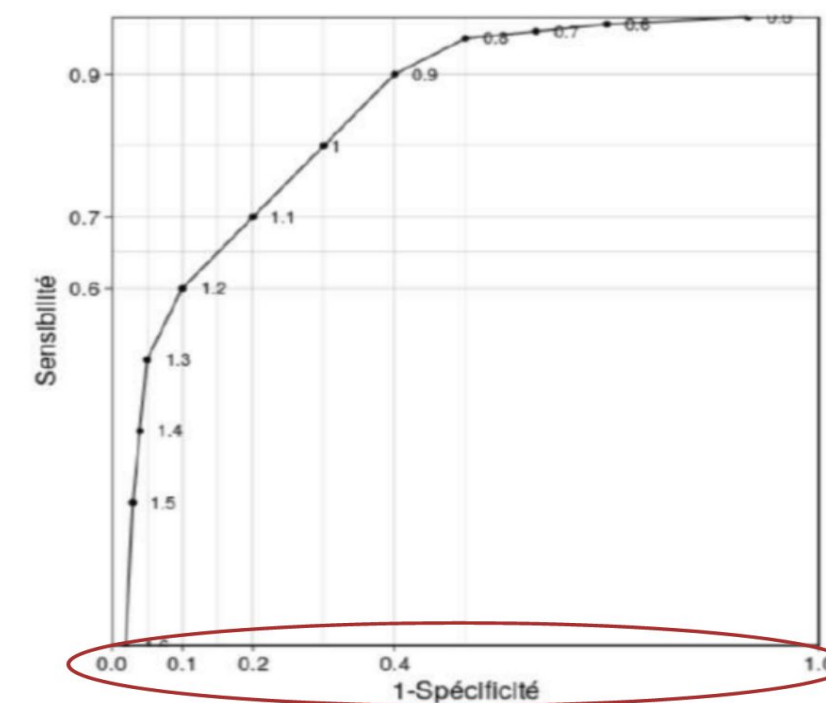
Lire une courbe ROC

Quand on recherche la probabilité d'avoir un test positif/négatif quand on est malade/sain :

Si on est malade on regarde en ordonnée :



Si on est sain on regarde en abscisse :



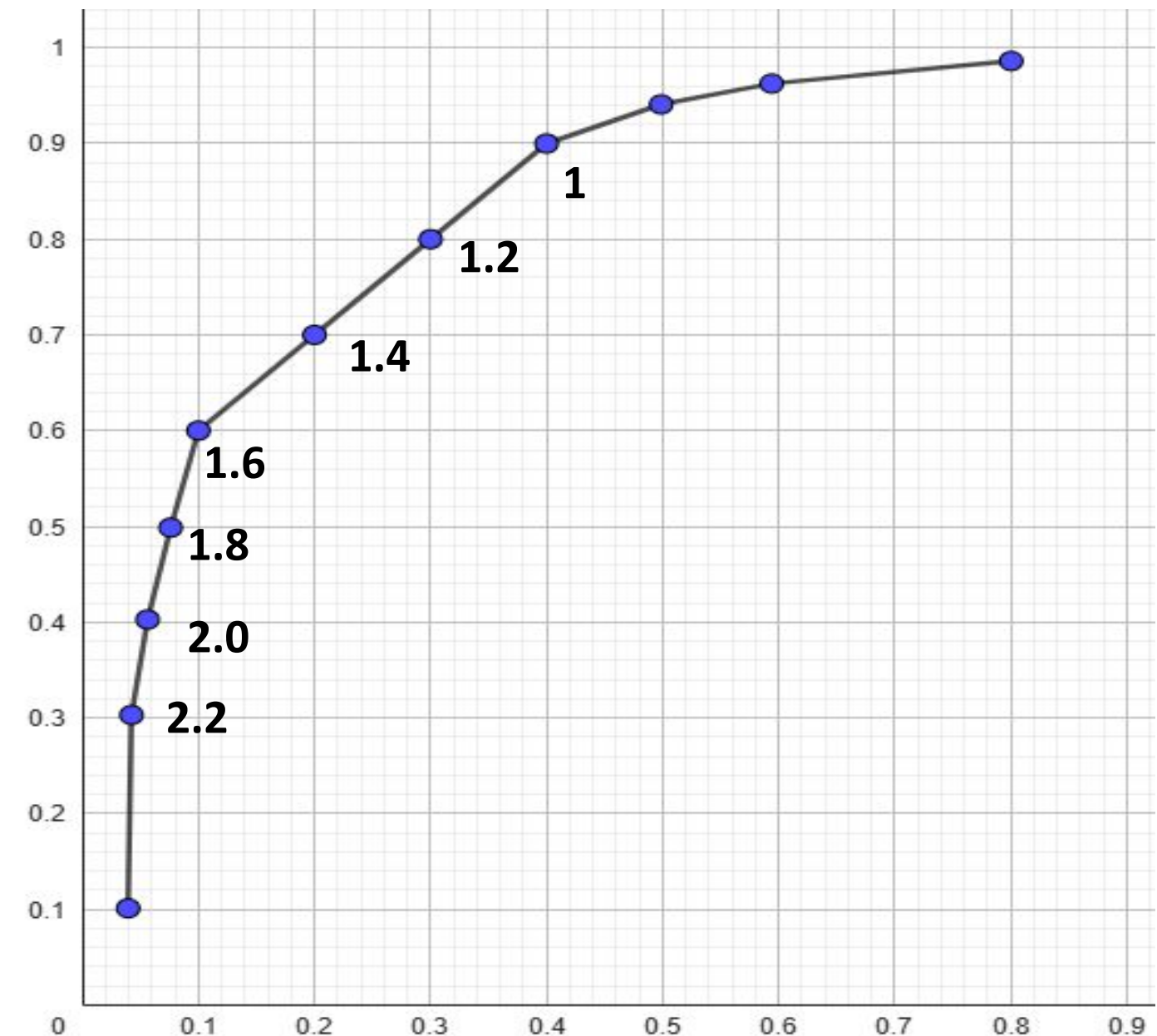
Si on cherche la probabilité d'avoir un test positif on prend la valeur lue

Si on cherche la probabilité d'avoir un test négatif on prend 1- la valeur lue

III - Courbe ROC

Que pensez-vous des assertions suivantes ?

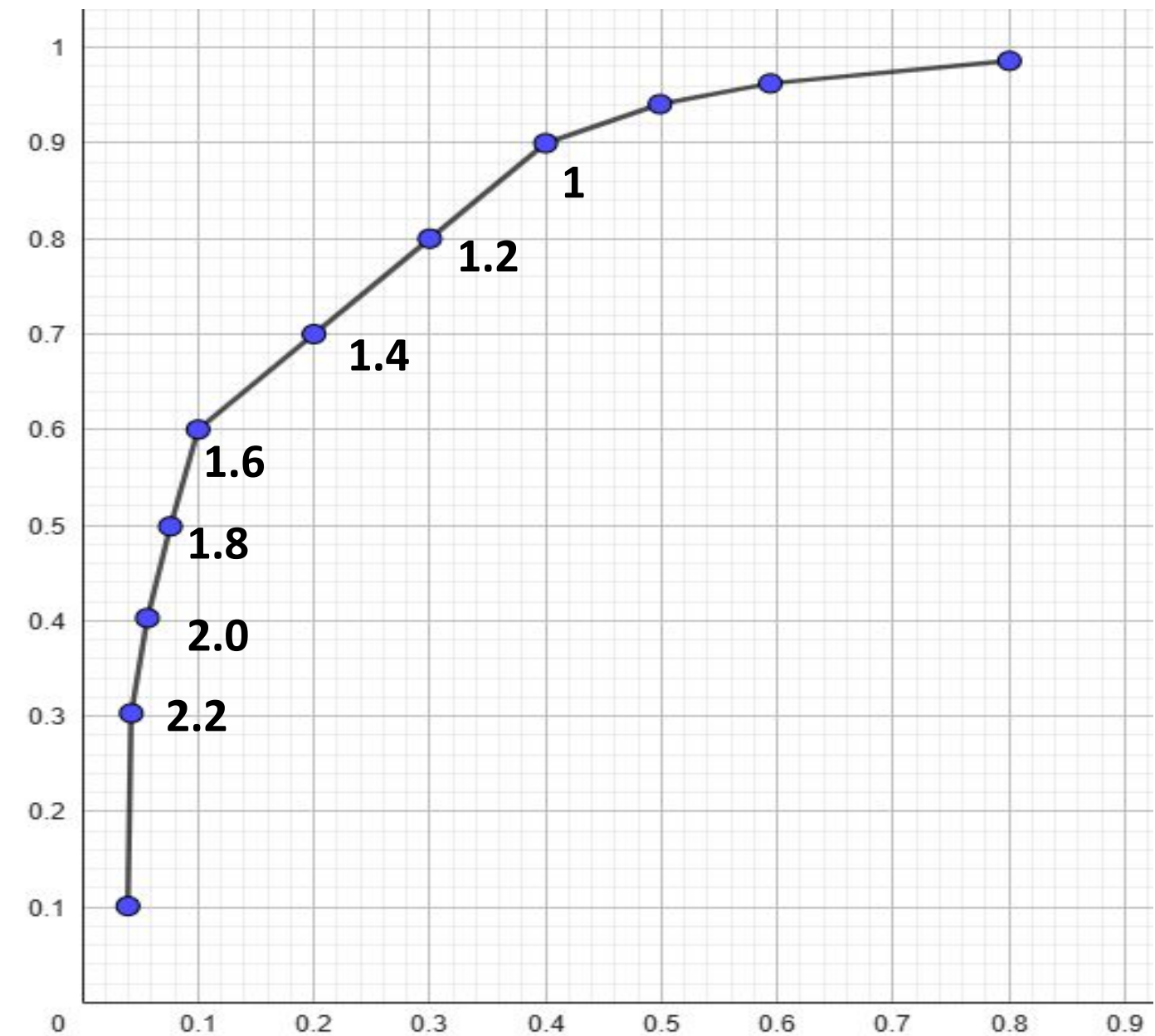
- A. Au seuil de dosage 1.6, la probabilité d'avoir un test positif chez les individus sains vaut 0,1
- B. Au seuil de dosage 1.6, la probabilité d'être malade si le test est négatif vaut 0,4
- C. Un test positif correspond à un dosage inférieur au seuil
- D. La valeur du dosage est plus importante chez les malades



III - Courbe ROC

Que pensez-vous des assertions suivantes ?

- A. Au seuil de dosage 1.6, la probabilité d'avoir un test positif chez les individus sains vaut 0,1
- B. Au seuil de dosage 1.6, la probabilité d'être malade si le test est négatif vaut 0,4
- C. Un test positif correspond à un dosage inférieur au seuil
- D. La valeur du dosage est plus importante chez les malades



III - Courbe ROC

A. Au seuil de dosage 1.6, la probabilité d'avoir un test positif chez les individus sains vaut 0,1

$$\rightarrow P(T+/M-) = 1 - P(T-/M-) = 1 - Sp = 0,1$$

B. Au seuil de dosage 1.6, la probabilité d'être malade si le test est négatif vaut 0,4

$$\rightarrow P(M+/T-) = 1 - P(M-/T-) = 1 - VPN.$$

VPN n'étant pas lisible sur le graphique on ne peut pas répondre à cette question

C. Un test positif correspond à un dosage inférieur au seuil

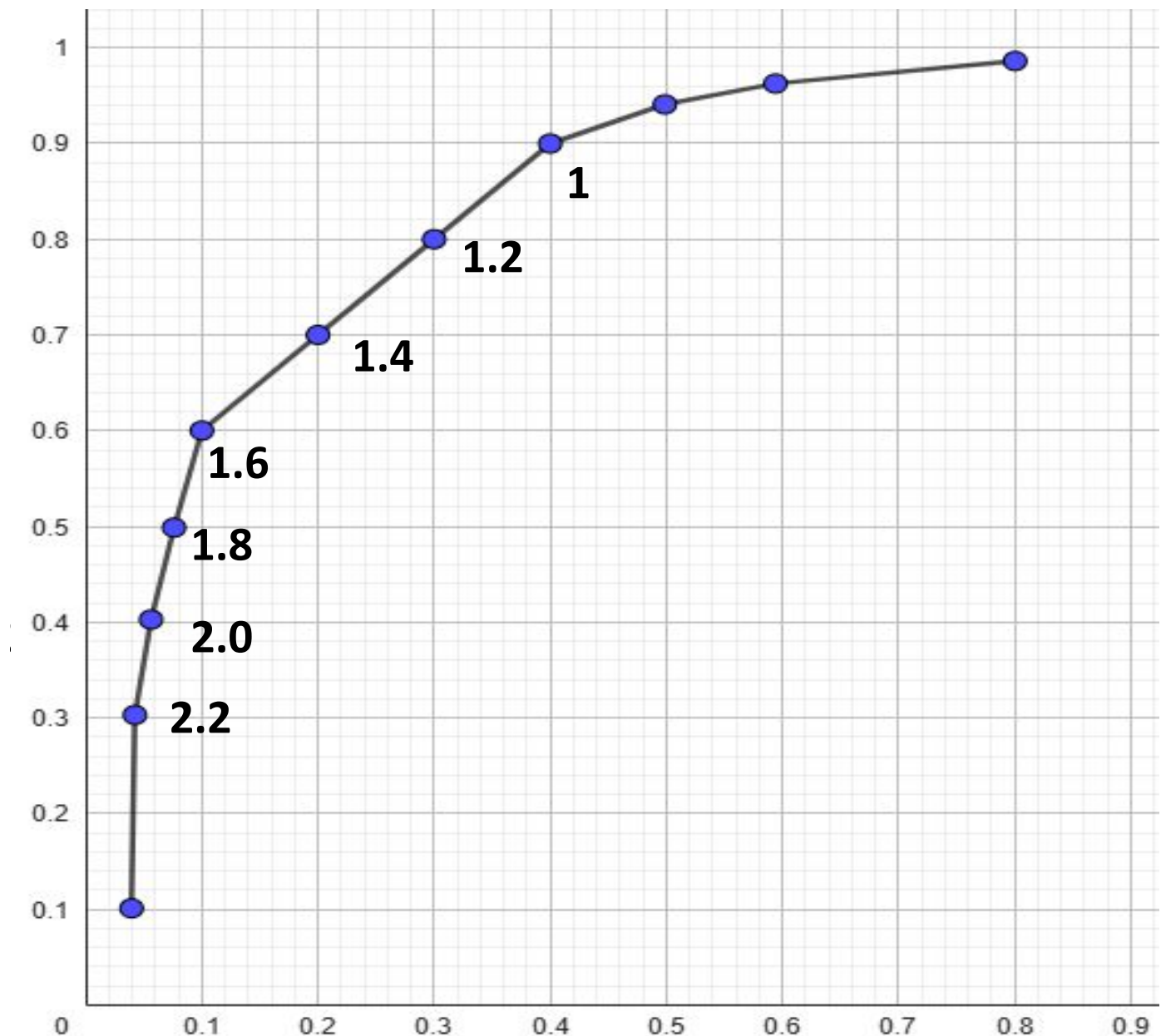
→ Par lecture graphique on voit que lorsque le seuil augmente, la sensibilité diminue. Donc un test positif correspond à un dosage supérieur au seuil.

D. La valeur du dosage est plus importante chez les malades

III - Courbe ROC

Suite

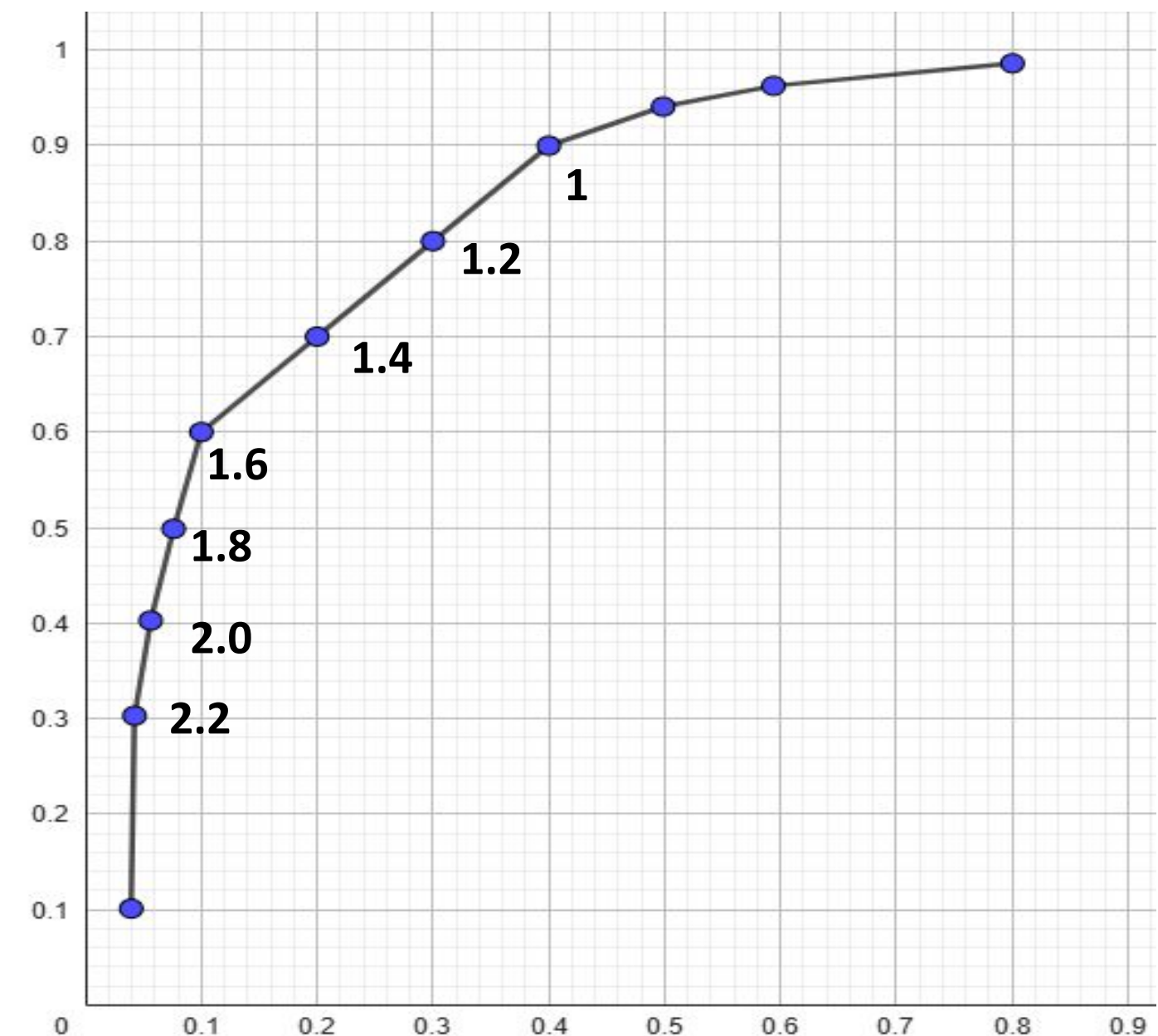
- A. Au seuil 1.6, l'indice de Youden vaut 0,7
- B. Au seuil 2.0, l'indice de Youden vaut 0,5
- C. Les données de la figure permettent de calculer la valeur prédictive négative pour chacun des seuils
- D. Les données de la figure permettent de calculer la probabilité d'avoir un test négatif pour chacun des seuils
- E. La probabilité d'être un vrai négatif diminue lorsque le seuil de dosage augmente



III - Courbe ROC

Suite

- A. Au seuil 1.6, l'indice de Youden vaut 0,7
- B. Au seuil 2.0, l'indice de Youden vaut 0,5
- C. Les données de la figure permettent de calculer la valeur prédictive négative pour chacun des seuils
- D. Les données de la figure permettent de calculer la probabilité d'avoir un test négatif pour chacun des seuils
- E. La probabilité d'être un vrai négatif diminue lorsque le seuil de dosage augmente



III - Courbe ROC

A. Au seuil 1.6, l'indice de Youden vaut 0,7

$$\rightarrow Se + Sp - 1 = 0,6 + 0,9 - 1 = 0,5$$

B. Au seuil 2.0, l'indice de Youden vaut 0,5

$$\rightarrow Se + Sp - 1 = 0,4 + 0,95 - 1 = 0,35$$

C. Les données de la figure permettent de calculer la valeur prédictive négative pour chacun des seuils

→ VPP et VPN incalculable avec uniquement une courbe ROC

D. Les données de la figure permettent de calculer la probabilité d'avoir un test négatif pour chacun des seuils

→ la courbe ROC nous permet uniquement d'accéder à la Se et la Sp pour chacun des seuils

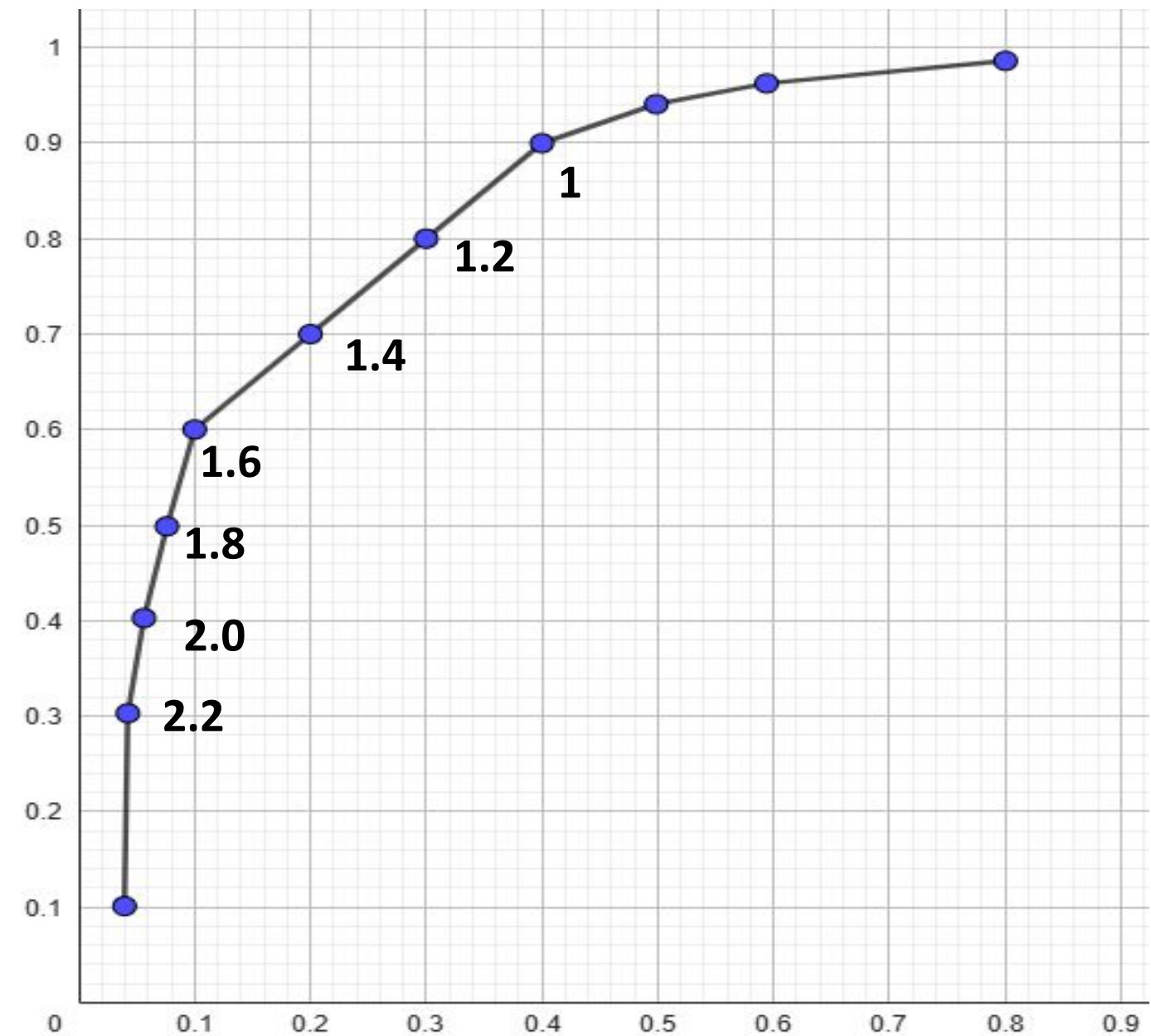
E. La probabilité d'être un vrai négatif diminue lorsque le seuil de dosage augmente

→ La Sp augmente quand le seuil augmente, soit la probabilité d'être un vrai négatif augmente lorsque le seuil de dosage augmente

III - Courbe ROC

Suite

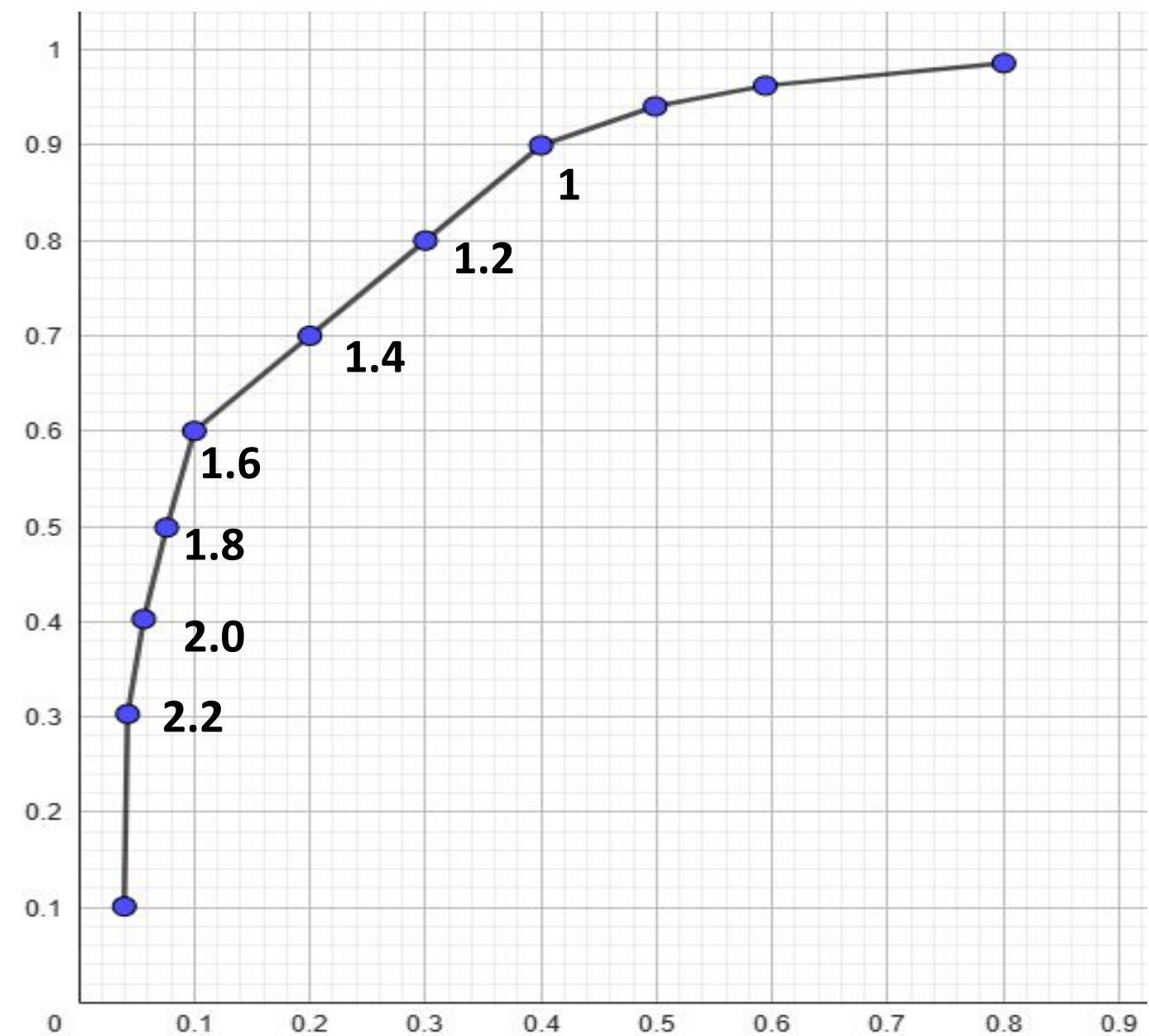
- A. Au seuil 1.6, le rapport de vraisemblance positif vaut 6
- B. Au seuil 1.6, le rapport de vraisemblance positif vaut $1/6$
- C. Au seuil 1.6, le rapport de vraisemblance négatif vaut $6/9$
- D. Au seuil 1.6, le rapport de vraisemblance négatif vaut $4/9$
- E. La contribution d'un test négatif est plus important au seuil 1.6 qu'au seuil 2.2



III - Courbe ROC

Suite

- A. Au seuil 1.6, le rapport de vraisemblance positif vaut 6
- B. Au seuil 1.6, le rapport de vraisemblance positif vaut $1/6$
- C. Au seuil 1.6, le rapport de vraisemblance négatif vaut $6/9$
- D. Au seuil 1.6, le rapport de vraisemblance négatif vaut $4/9$
- E. La contribution d'un test négatif est plus important au seuil 1.6 qu'au seuil 2.2



III - Courbe ROC

A. Au seuil 1.6, le rapport de vraisemblance positif vaut 6

$$\rightarrow L = Se/1-Sp = 0,6/0,1 = 6$$

B. Au seuil 1.6, le rapport de vraisemblance positif vaut 1/6

C. Au seuil 1.6, le rapport de vraisemblance négatif vaut 6/9

$$\rightarrow \lambda = 1-Se/Sp = 0,4/0,9 = 4/9$$

D. Au seuil 1.6, le rapport de vraisemblance négatif vaut 4/9

E. La contribution d'un test négatif est plus important au seuil 1.6 qu'au seuil 2.2

→ Pour avoir la contribution la plus importante possible:

- La vraisemblance positive doit être la + grande possible
- La vraisemblance négative doit être la - grande possible

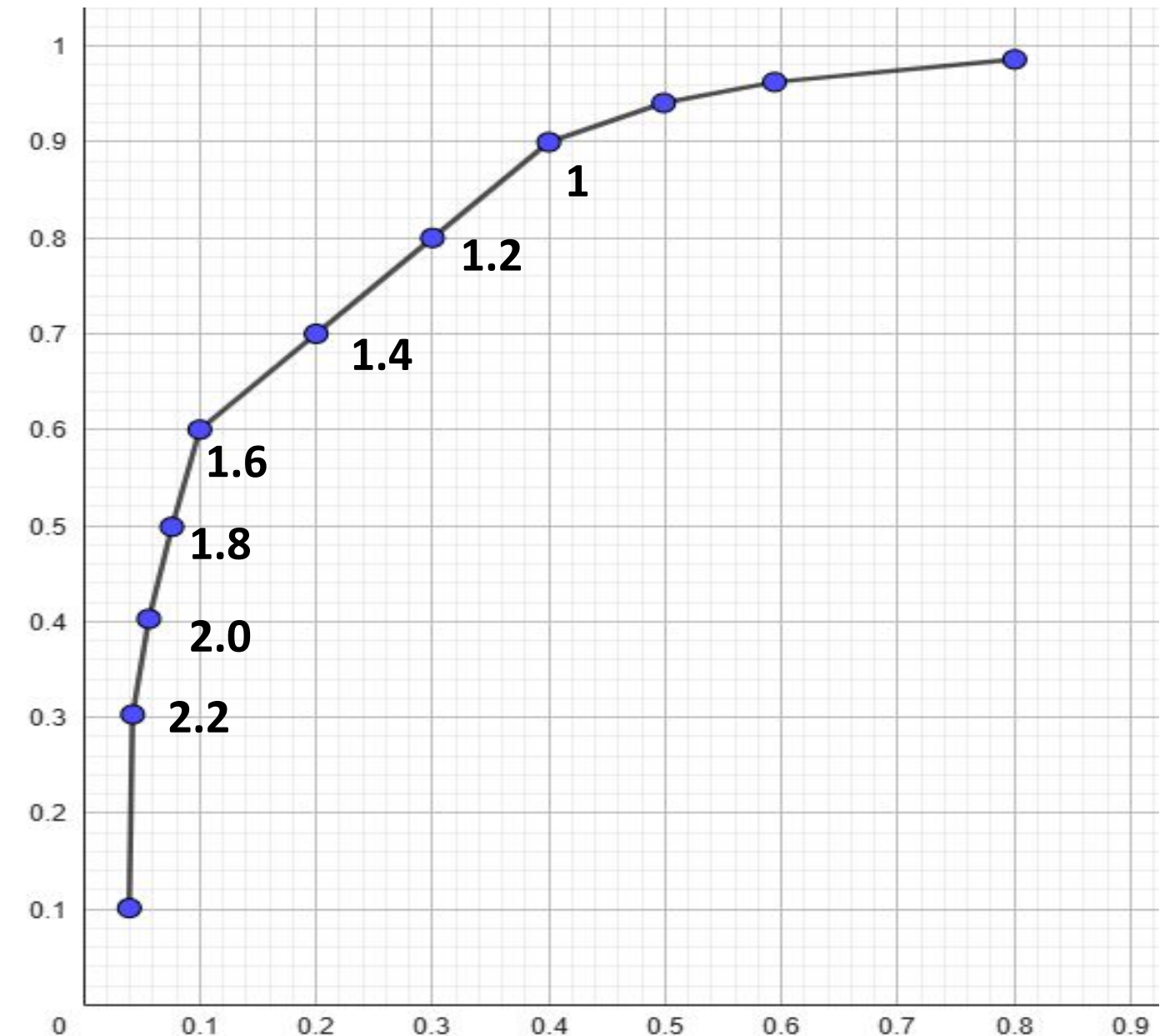
➤ Au seuil 2.2, $\lambda = 1-Se/Sp = 0,7/0,97 \approx 0,7 > 4/9$ du seuil 1,6

III - Courbe ROC

On dispose pour la question suivante d'informations complémentaires sur l'échantillon représentatif de la population cible de 1000 sujets qui a servi à l'évaluation de ce test diagnostic: dans la population cible, la prévalence de la maladie vaut 10%

Pour un seuil de 1.6, que pensez vous des propositions suivantes ? Dans cet échantillon:

- A. On compte 810 vrais négatifs
- B. On compte 900 vrais négatifs
- C. On compte 60 vrais positifs
- D. On compte 40 faux négatifs
- E. On compte 40 faux positifs

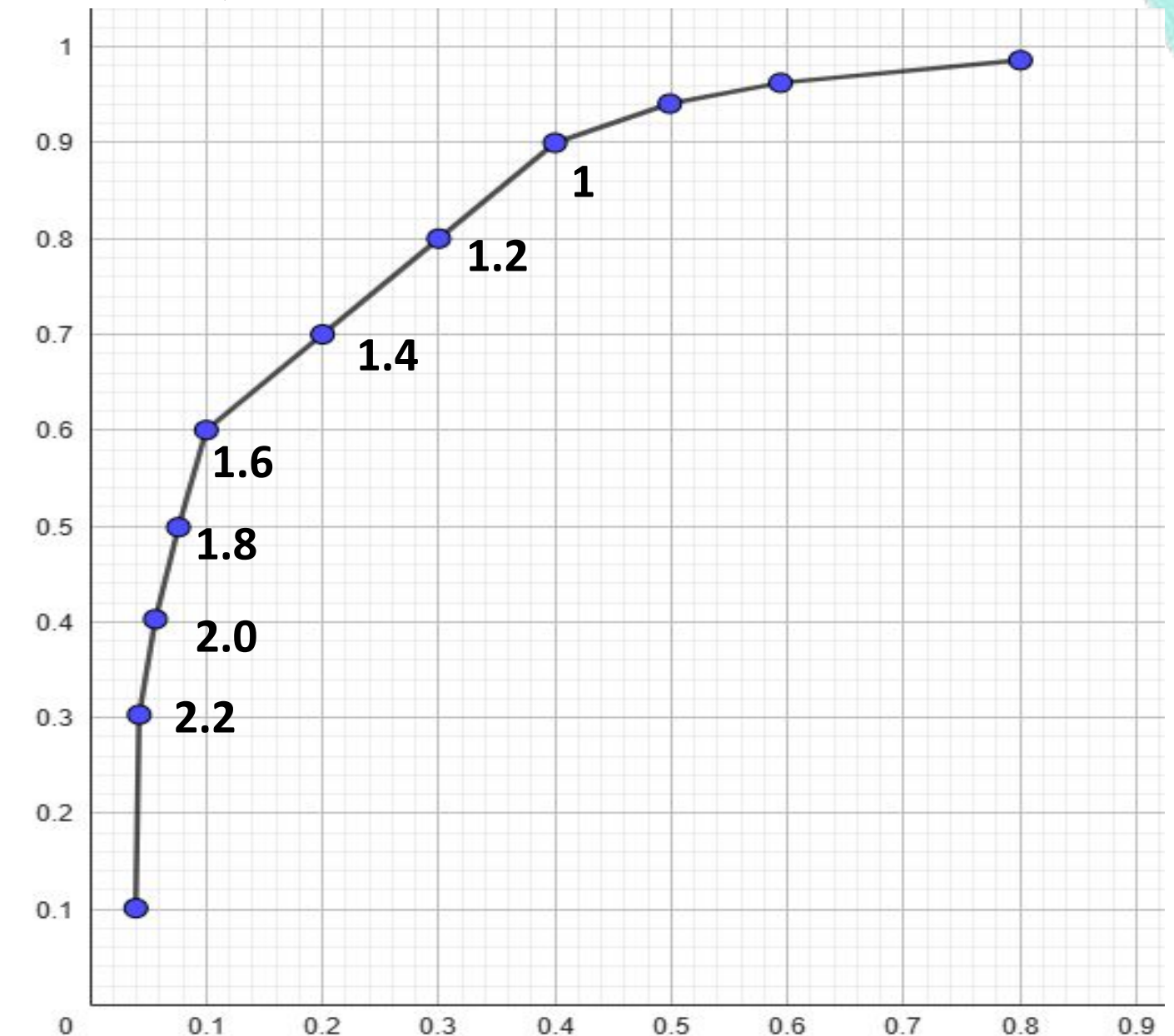


III - Courbe ROC

On dispose pour la question suivante d'informations complémentaires sur l'échantillon représentatif de la population cible de 1000 sujets qui a servi à l'évaluation de ce test diagnostic: dans la population cible, la prévalence de la maladie vaut 10%

Pour un seuil de 1.6, que pensez vous des propositions suivantes ? Dans cet échantillon:

- A. On compte 810 vrais négatifs
- B. On compte 900 vrais négatifs
- C. On compte 60 vrais positifs
- D. On compte 40 faux négatifs
- E. On compte 40 faux positifs



III - Courbe ROC

Pour un seuil de 1.6, que pensez vous des propositions suivantes ?
Dans cet échantillon:

- A. On compte 810 vrais négatifs
- B. On compte 900 vrais négatifs
- C. On compte 60 vrais positifs
- D. On compte 40 faux négatifs
- E. On compte 40 faux positifs

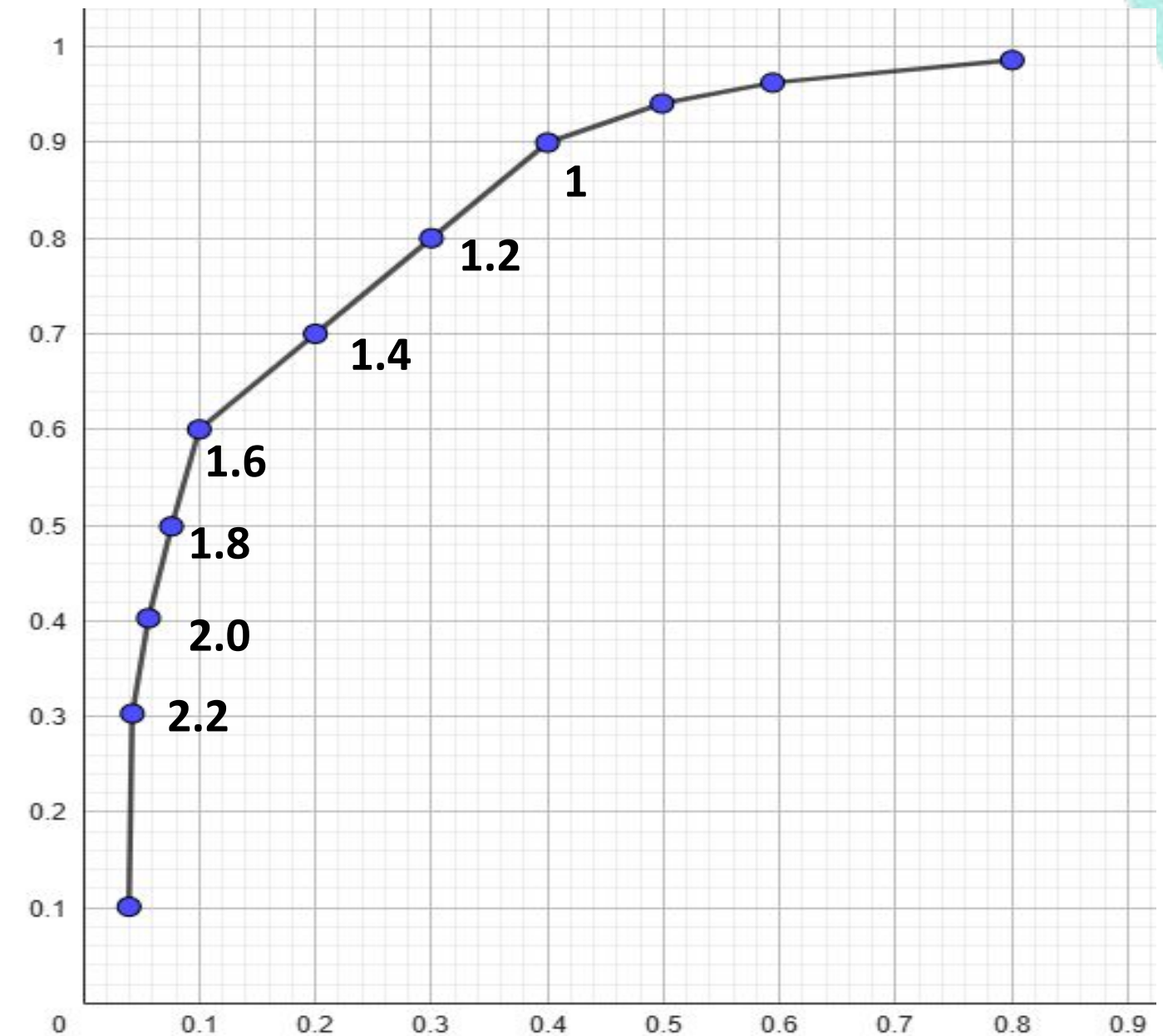
$$VP = Se \times M+ = 0,6 \times 100 = 60$$
$$VN = Sp \times M- = 0,9 \times 900 = 810$$

	M+	M-	Totaux
T+	VP=60	FP=90	150
T-	FN=40	VN=810	850
Totaux	10% de 1000 soit 100	900	N=1000

III - Courbe ROC

Pour un seuil de 1.6, que pensez vous des propositions suivantes ?

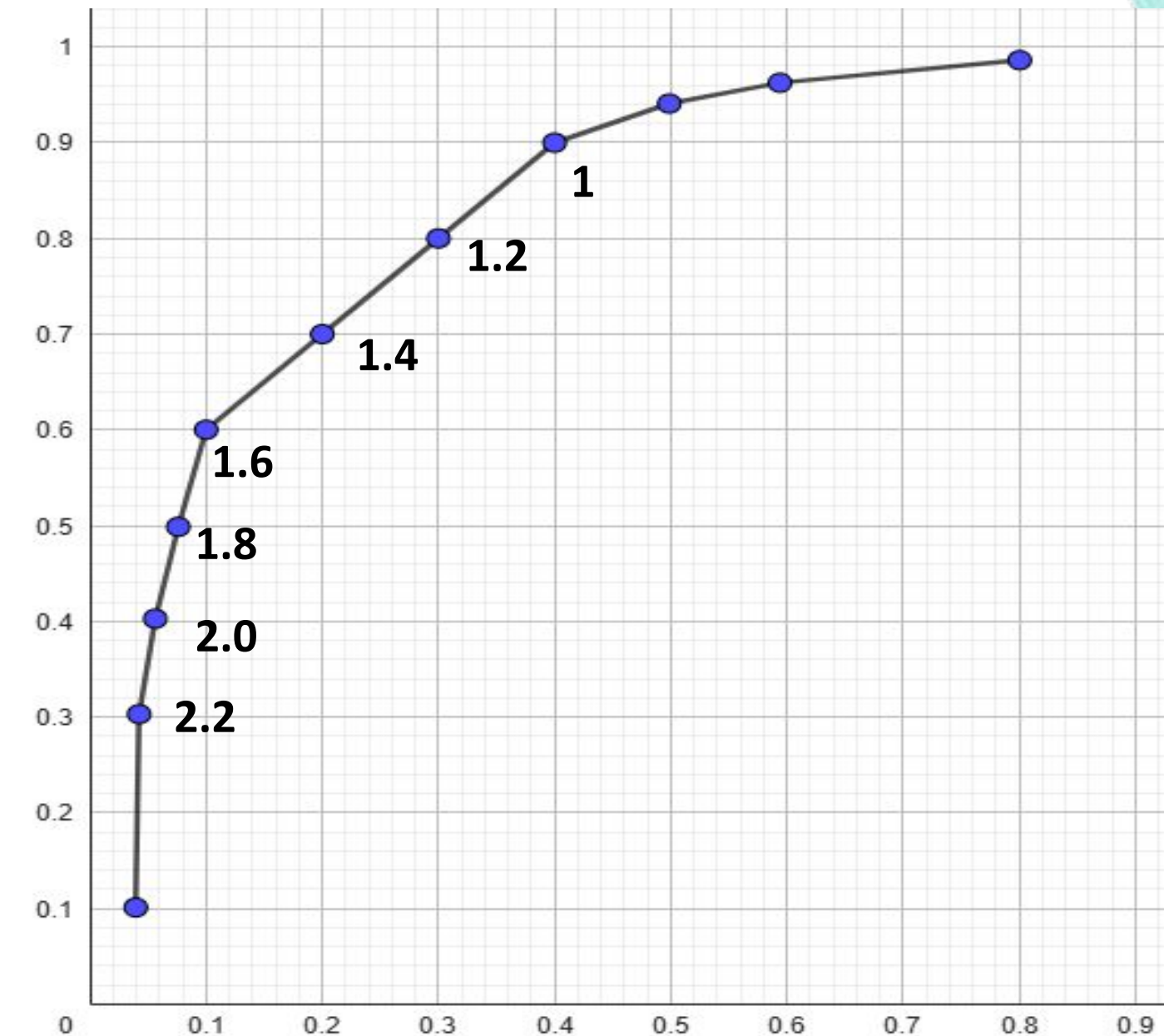
- A. La valeur prédictive positive vaut 60/150
- B. La valeur prédictive positive vaut 60/100
- C. La valeur prédictive négative vaut 810/850
- D. La valeur prédictive négative vaut 90/900
- E. La probabilité d'avoir un test positif vaut 0,15



III - Courbe ROC

Pour un seuil de 1.6, que pensez vous des propositions suivantes ?

- A. La valeur prédictive positive vaut 60/150
- B. La valeur prédictive positive vaut 60/100
- C. La valeur prédictive négative vaut 810/850
- D. La valeur prédictive négative vaut 90/900
- E. La probabilité d'avoir un test positif vaut 0,15



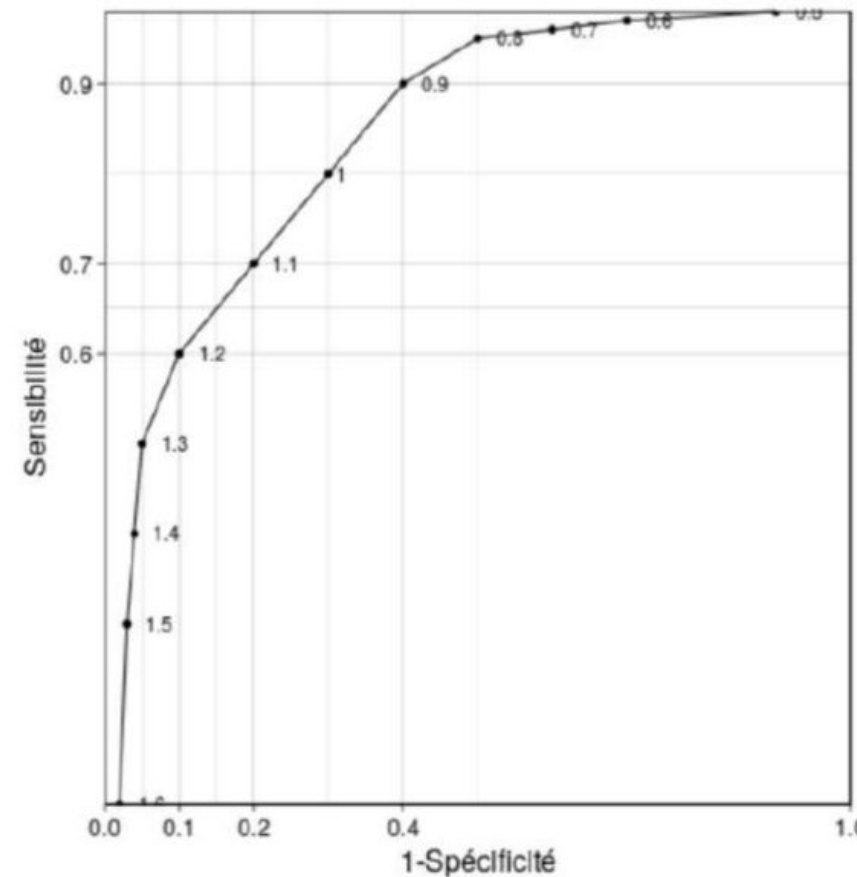
III - Courbe ROC

Pour un seuil de 1.6, que pensez vous des propositions suivantes ?

- A. La valeur prédictive positive vaut 60/150
→ $VPP = P(M+/T+) = VP/T+ = 60/150$
- B. La valeur prédictive positive vaut 60/100
- C. La valeur prédictive négative vaut 810/850
→ $VPN = P(M-/T-) = VN/T- = 810/850$
- D. La valeur prédictive négative vaut 90/900
- E. La probabilité d'avoir un test positif vaut 0,15
→ $P(T+) = T+/N = 150/1000 = 0,15$

III - Courbe ROC

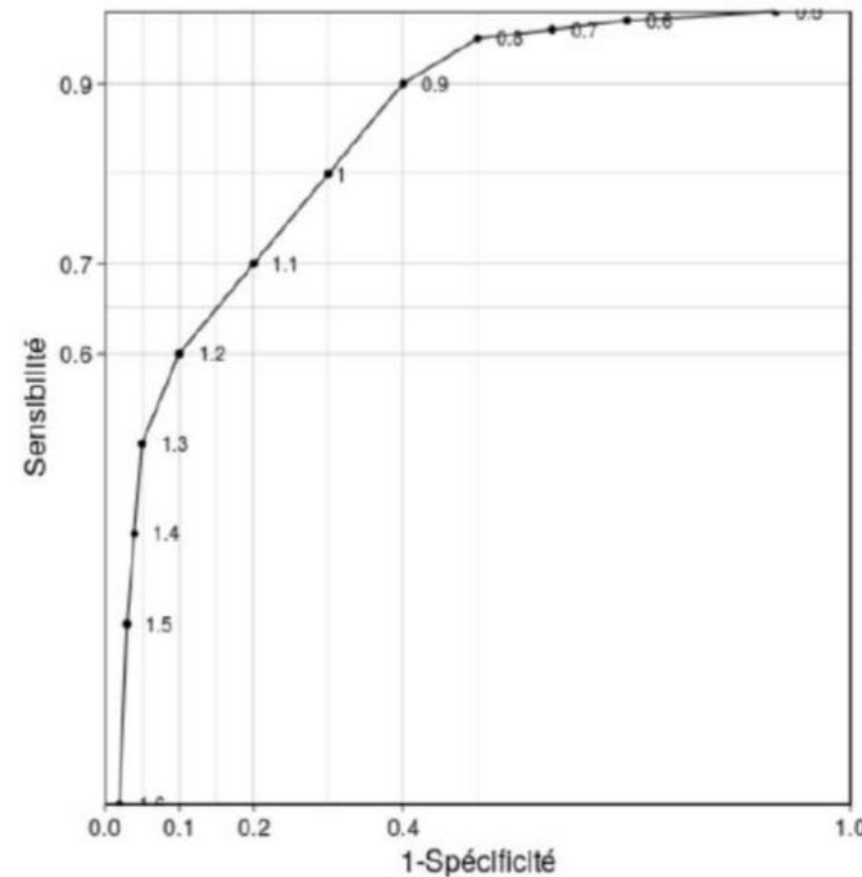
L'évaluation d'un test diagnostique basé sur un dosage biologique donne la courbe ROC suivante. Un test positif est censé détecter la maladie. Sur chaque point figure la valeur du seuil étudié.



- A. Un test négatif correspond à un dosage inférieur au seuil
- B. La sensibilité augmente quand le seuil du dosage augmente
- C. Au seuil de dosage 0,9, la probabilité d'avoir un test positif si on est malade est de 0,1
- D. Au seuil de dosage de 0,9, la probabilité d'avoir un test positif si on est sain est de 0,4
- E. Au seuil de dosage 1,1, la probabilité d'avoir un test négatif si on est malade est de 0,3

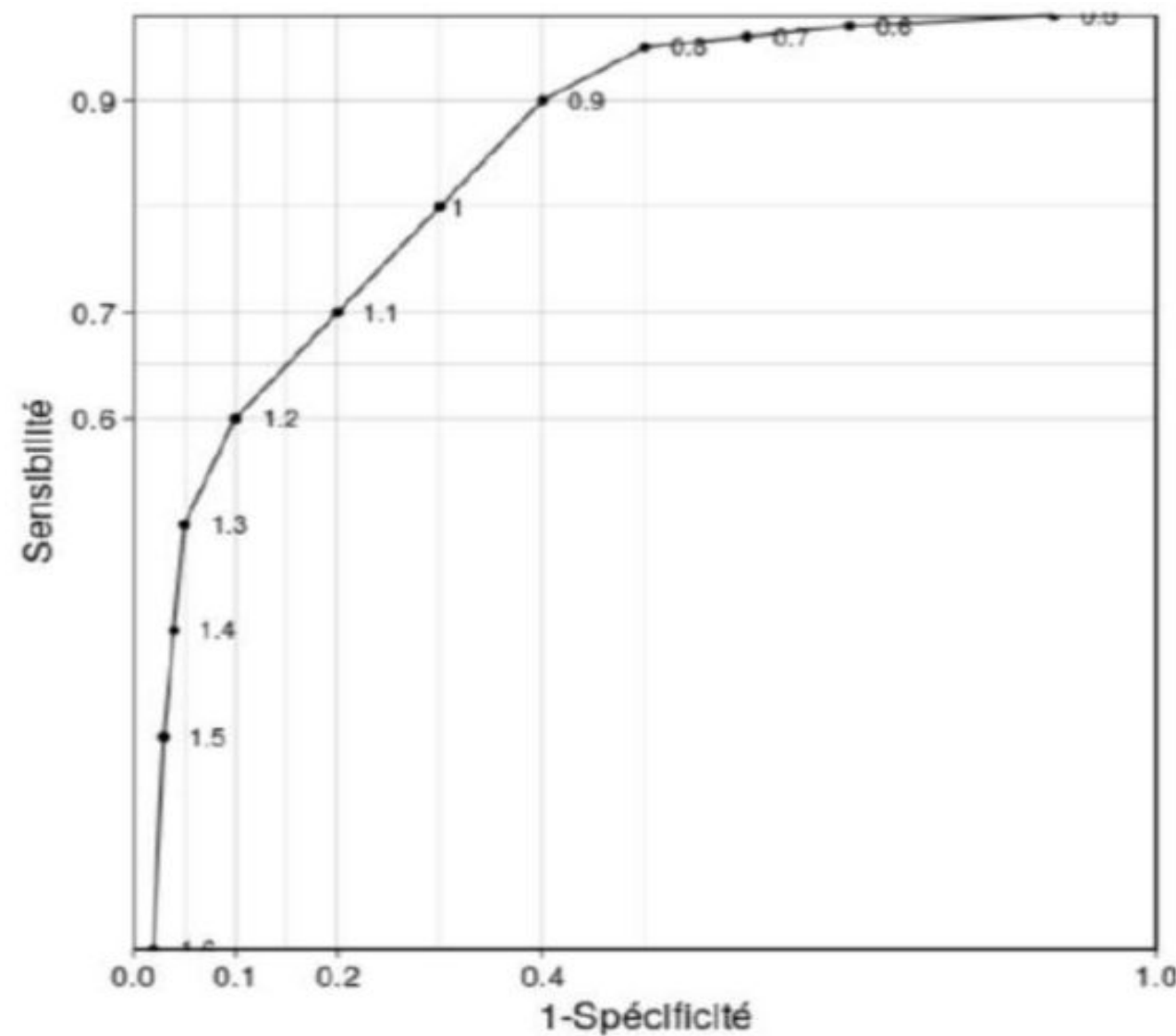
III - Courbe ROC

L'évaluation d'un test diagnostique basé sur un dosage biologique donne la courbe ROC suivante. Un test positif est censé détecter la maladie. Sur chaque point figure la valeur du seuil étudié.



- A. Un test négatif correspond à un dosage inférieur au seuil
- B. La sensibilité augmente quand le seuil du dosage augmente
- C. Au seuil de dosage 0,9, la probabilité d'avoir un test positif si on est malade est de 0,1
- D. Au seuil de dosage de 0,9, la probabilité d'avoir un test positif si on est sain est de 0,4
- E. Au seuil de dosage 1,1, la probabilité d'avoir un test négatif si on est malade est de 0,3

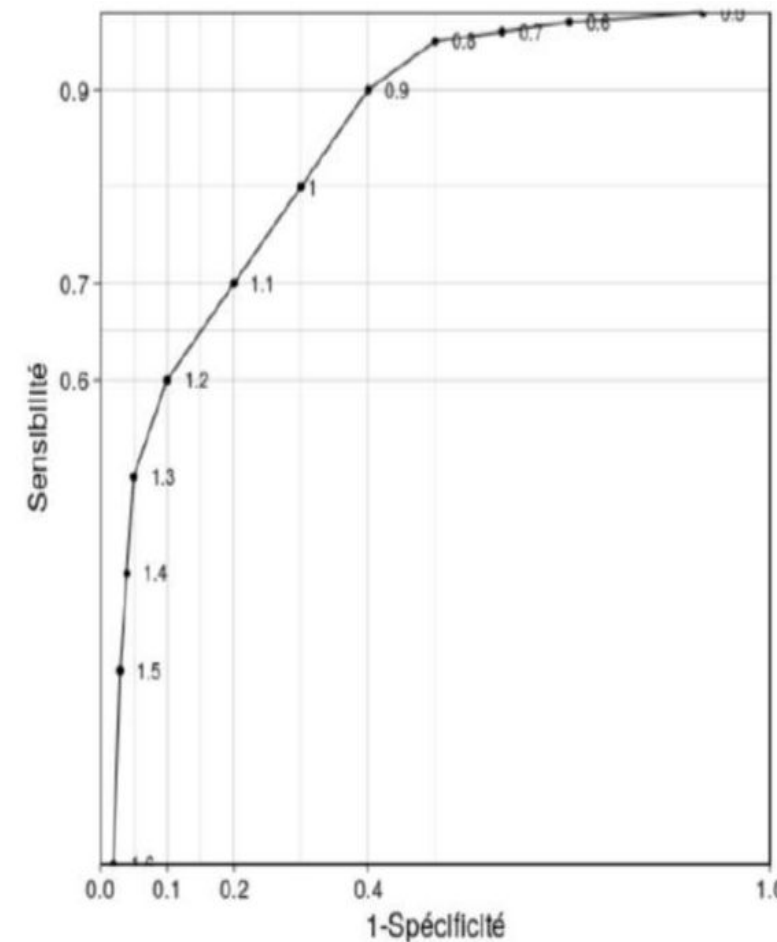
III - Courbe ROC



- A. Un test négatif correspond à un dosage inférieur au seuil
→ Car quand le seuil diminue, Se augmente.
- B. La sensibilité augmente quand le seuil du dosage augmente.
→ Quand le seuil du dosage diminue
- C. Au seuil de dosage 0,9, la probabilité d'avoir un test positif si on est malade est de 0,1
→ Il faut lire la sensibilité: 0,9.
- D. Au seuil de dosage de 0,9, la probabilité d'avoir un test positif si on est sain est de 0,4
→ Ici on cherche 1 - Sp soit 0,4
- E. Au seuil de dosage 1,1, la probabilité d'avoir un test négatif si on est malade est de 0,3

III - Courbe ROC

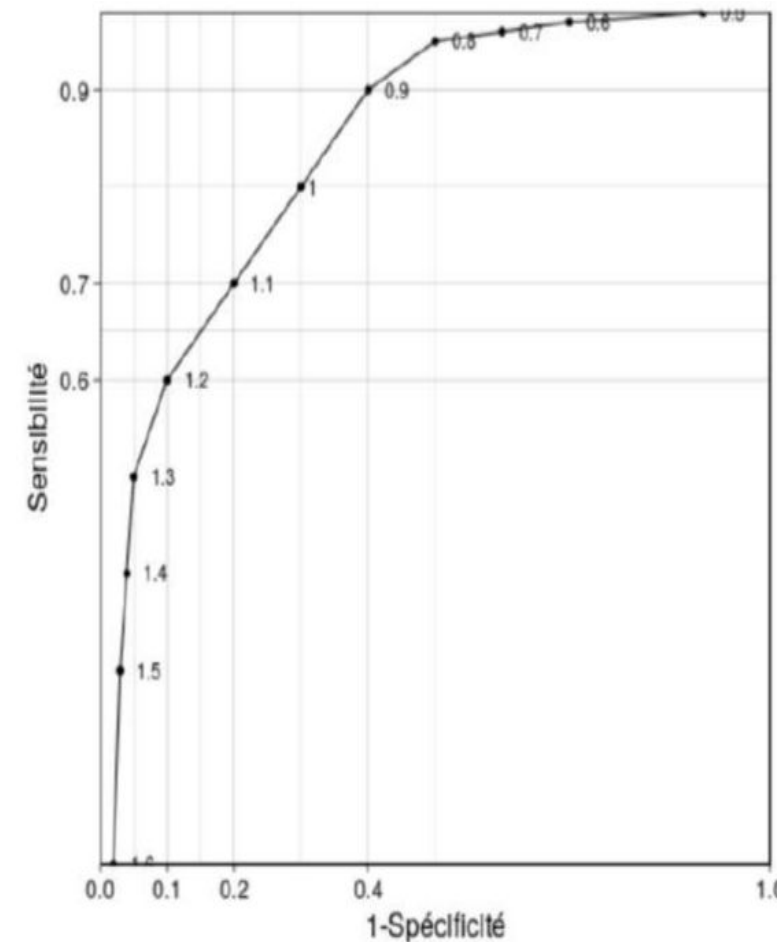
SUITE :



- A. Les données de la figure permettent de calculer la probabilité d'avoir un test négatif pour chacun des seuils.
- B. Les données de la figure permettent de calculer la valeur prédictive positive pour chacun des seuils.
- C. La probabilité d'être un faux négatif augmente quand le seuil du dosage augmente.
- D. Pour une spécificité de 0,6 l'indice de Youden vaut 0,4.
- E. Au seuil de dosage 1,2 l'indice de Youden vaut 0,1.

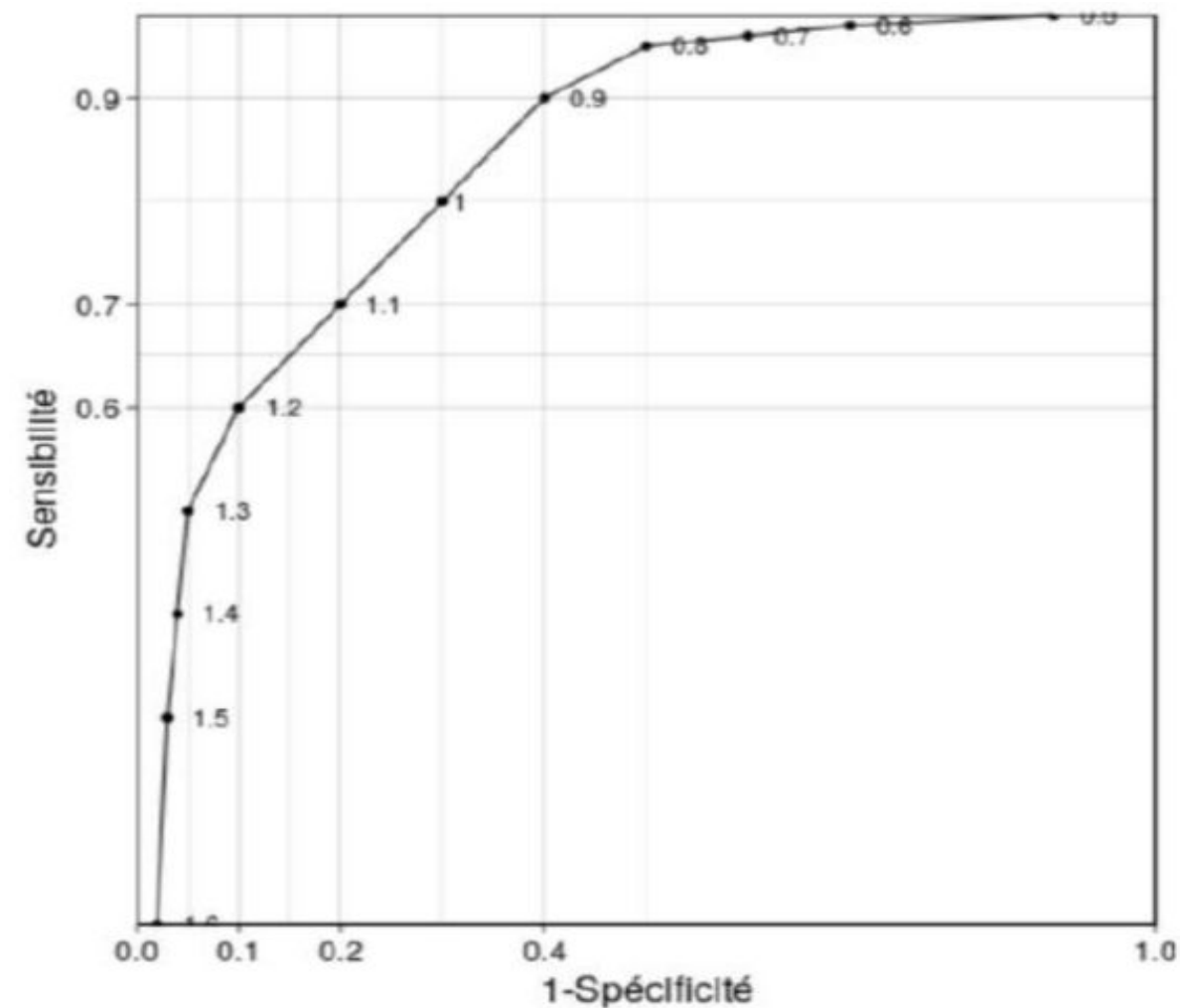
III - Courbe ROC

SUITE :



- A. Les données de la figure permettent de calculer la probabilité d'avoir un test négatif pour chacun des seuils.
- B. Les données de la figure permettent de calculer la valeur prédictive positive pour chacun des seuils.
- C. La probabilité d'être un faux négatif augmente quand le seuil du dosage augmente.
- D. Pour une spécificité de 0,6 l'indice de Youden vaut 0,4.
- E. Au seuil de dosage 1,2 l'indice de Youden vaut 0,1.

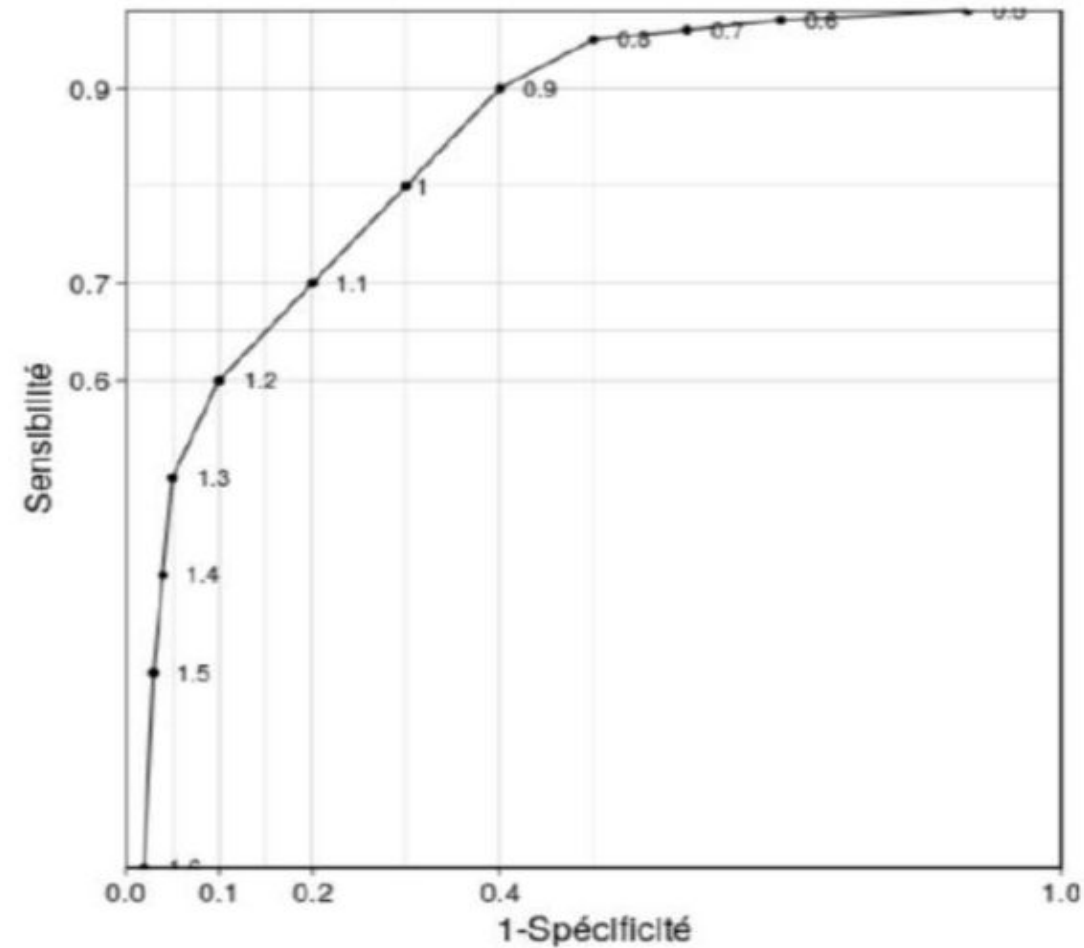
III - Courbe ROC



- A. Les données de la figure permettent de calculer la probabilité d'avoir un test négatif pour chacun des seuils.**
→ il manquerait la prévalence
- B. Les données de la figure permettent de calculer la valeur prédictive positive pour chacun des seuils.**
→ Idem
- C. La probabilité d'être un faux négatif augmente quand le seuil du dosage augmente.**
- D. Pour une spécificité de 0,6 l'indice de Youden vaut 0,4.**
→ $Spe + Se - 1 = 0,6 + 0,9 - 1 = 0,5$
- E. Au seuil de dosage 1,2 l'inde de Youden vaut 0,1.**
→ $Spe + Se - 1 = 0,9 + 0,6 - 1 = 0,5$

III - Courbe ROC

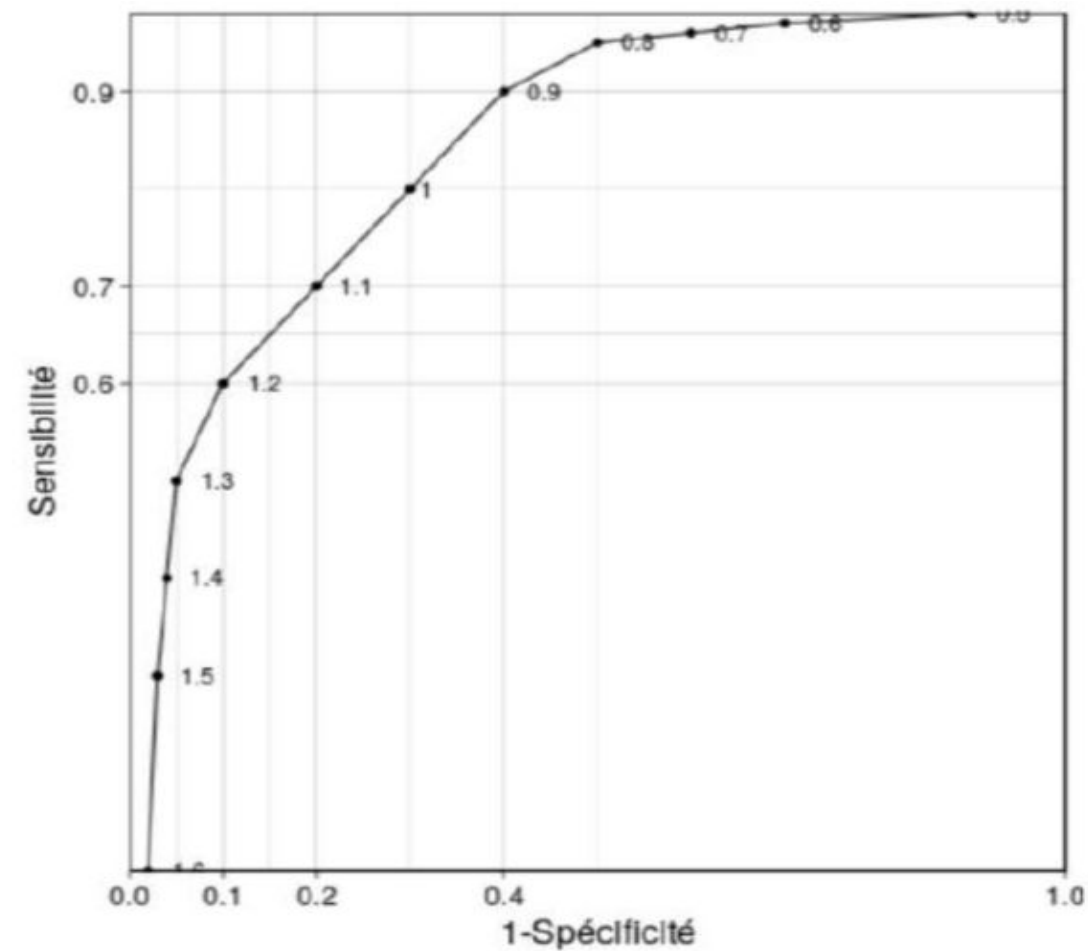
SUITE :



- A. Au seuil de dosage 1,1, le rapport de vraisemblance positive vaut $\frac{7}{8}$
- B. Au seuil de dosage 1,1 le rapport de vraisemblance positive vaut $\frac{7}{2}$
- C. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{7}{8}$
- D. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{3}{8}$
- E. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{5}{8}$

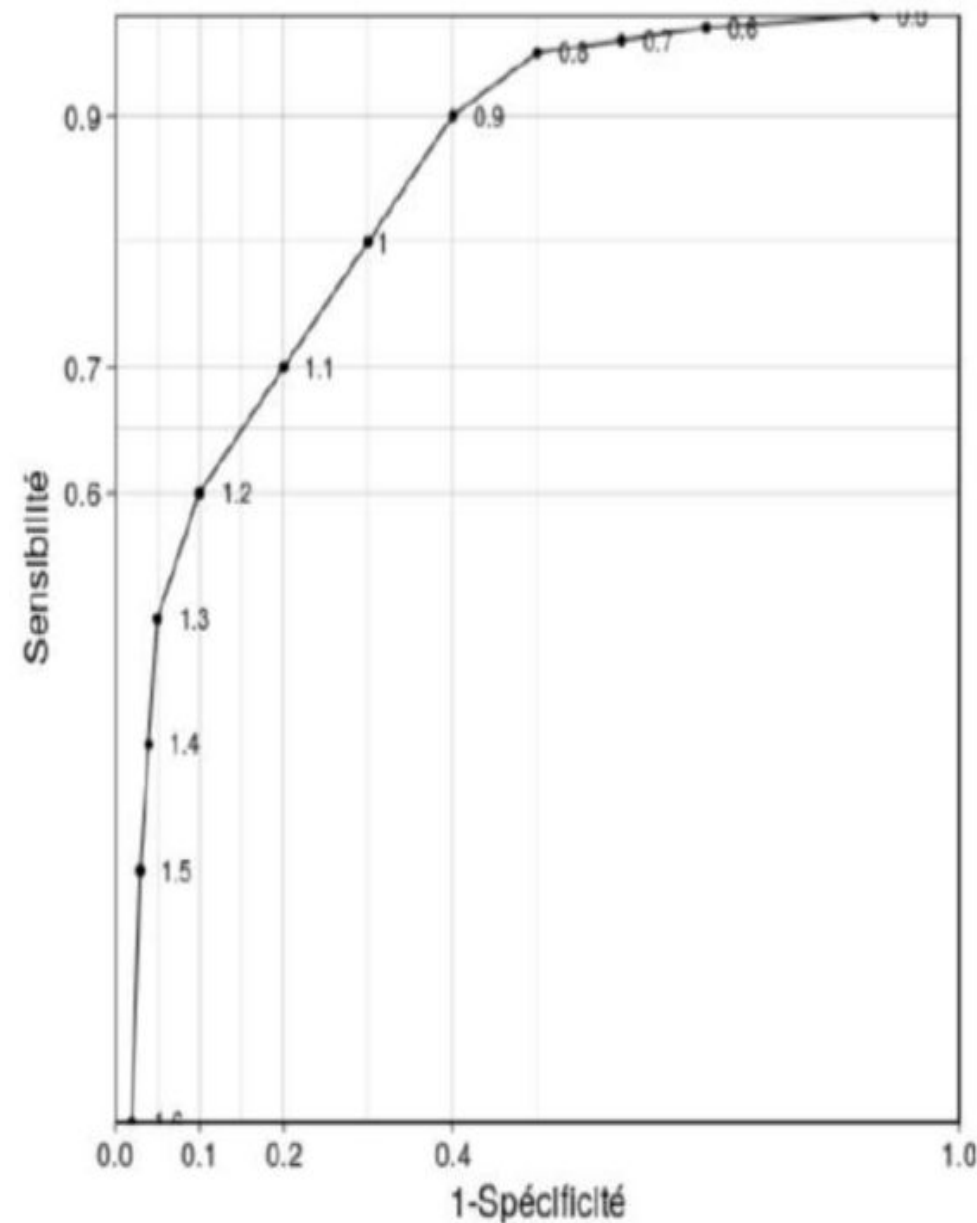
III - Courbe ROC

SUITE :



- A. Au seuil de dosage 1,1, le rapport de vraisemblance positive vaut $\frac{7}{8}$
- B. Au seuil de dosage 1,1 le rapport de vraisemblance positive vaut $\frac{7}{2}$
- C. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{7}{8}$
- D. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{3}{8}$
- E. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{5}{8}$

III - Courbe ROC



- A. Au seuil de dosage 1,1, le rapport de vraisemblance positive vaut $\frac{7}{8}$.**
- B. Au seuil de dosage 1,1 le rapport de vraisemblance positive vaut $\frac{7}{2}$.**
→ Au seuil 1,1 : $L = \text{Se} / (1 - \text{spe}) = 0,7 / 0,2 = 7/2$
- C. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{7}{8}$.**
- D. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{3}{8}$;**
- E. Au seuil de dosage 1,1, le rapport de vraisemblance négative vaut $\frac{5}{8}$.**
→ Au seuil 1,1 : $\lambda = (1 - \text{Se}) / \text{Spe} = 0,3 / 0,8 = 3/8$

IV - Annales

Le tableau ci-dessous vous donne les effectifs observés dans le cadre de l'étude d'un test biologique chez 166 sujets.

Cochez pour chaque affirmation si elle est vraie ou fausse.

	Malades	Non malades
Test positif	60	18
Test négatif	30	58

- A. Il s'agit d'un tableau de contingence
- B. La valeur 118 correspond à un des effectifs marginaux que l'on peut calculer ici
- C. La fréquence observée de 60/90 ainsi que la fréquence observée de 60/78 peuvent être associées aux vrais positifs
- D. La sensibilité et la valeur prédictive positive auront le même numérateur
- E. Dans le calcul de la spécificité, le numérateur correspond aux vrais négatifs.

IV - Annales

Le tableau ci-dessous vous donne les effectifs observés dans le cadre de l'étude d'un test biologique chez 166 sujets.

Cochez pour chaque affirmation si elle est vraie ou fausse.

	Malades	Non malades
Test positif	60	18
Test négatif	30	58

- A. Il s'agit d'un tableau de contingence
- B. La valeur 118 correspond à un des effectifs marginaux que l'on peut calculer ici
- C. La fréquence observée de 60/90 ainsi que la fréquence observée de 60/78 peuvent être associées aux vrais positifs
- D. La sensibilité et la valeur prédictive positive auront le même numérateur
- E. Dans le calcul de la spécificité, le numérateur correspondra aux vrais négatifs.

IV - Annales

Le tableau ci-dessous vous donne les effectifs observés dans le cadre de l'étude d'un test biologique chez 166 sujets.

Cochez pour chaque affirmation si elle est vraie ou fausse.

	Malades	Non malades
Test positif	60	18
Test négatif	30	58

A. Il s'agit d'un tableau de contingence.

B. La valeur 118 correspond à un des effectifs marginaux que l'on peut calculer ici

→ Marginaux = totaux. 118 ne correspond à aucun effectif total en ligne ou en colonne, cela correspond au nombre de diagnostics corrects (VP + VN).

C. La fréquence observée de 60/90 ainsi que la fréquence observée de 60/78 peuvent être associées aux vrais positifs.

D. La sensibilité et la valeur prédictive positive auront le même numérateur

→ $Se = VP / (VP + FN)$ et $VPP = VP / (VP + FP)$

E. Dans le calcul de la spécificité, le numérateur correspondra aux vrais négatifs.

→ $Sp = VN / (VN + FP)$

IV - Annales

On réalise une étude dans un échantillon de 200 sujets, dont 100 malades pour connaître l'intérêt d'un test diagnostique dans une maladie M. Une partie des résultats figure dans le tableau suivant

	Test positif	Test négatif
Malades	80	?
Non malades	?	?

On sait également que 100 tests sont positifs. L'échantillon est représentatif de la population. Que pensez vous des affirmations suivantes ?

- A. La sensibilité du test vaut 80/100
- B. La spécificité du test vaut 0,8
- C. La valeur prédictive positive du test vaut 8/10
- D. La valeur prédictive négative du test vaut 80/100
- E. La probabilité que le test donne un résultat correct vaut 80/100

IV - Annales

On réalise une étude dans un échantillon de 200 sujets, dont 100 malades pour connaître l'intérêt d'un test diagnostique dans une maladie M. Une partie des résultats figure dans le tableau suivant

	Test positif	Test négatif
Malades	80	?
Non malades	?	?

On sait également que 100 tests sont positifs. L'échantillon est représentatif de la population. Que pensez vous des affirmations suivantes ?

- A. La sensibilité du test vaut 80/100
- B. La spécificité du test vaut 0,8
- C. La valeur prédictive positive du test vaut 8/10
- D. La valeur prédictive négative du test vaut 80/100
- E. La probabilité que le test donne un résultat correct vaut 80/100

IV - Annales

A. La sensibilité du test vaut 80/100.

$$Se = VP/(VP+FN) = 80/(80+20)$$

B. La spécificité du test vaut 0,8.

$$Sp = VN/(VN+FP) = 80/(80+20)$$

C. La valeur prédictive positive du test vaut 8/10.

$$VPP = VP/(VP+FP) = 80/(80+20)$$

D. La valeur prédictive négative du test vaut 80/100.

$$VPN = VN/(VN+FN) = 80/(80+20)$$

E. La probabilité que le test donne un résultat correct vaut 80/100.

Il est correct s'il est positif pour un malade et négatif pour un non malade :
 $(80+80)/200$

	Test positif	Test négatif
Malades	80	20
Non malades	20	80

IV - Annales

Le modèle de régression linéaire simple est défini de la façon suivante avec α et β comme paramètres du modèle :

$$Y = \beta + \alpha X + \varepsilon$$

Dans cette définition

- A. Y est un prédicteur
- B. α s'exprime avec la même unité que celle associée à X
- C. X et Y sont appariées
- D. X et Y correspondent aux inconnues du modèle
- E. ε s'exprime avec la même unité que celle associée à Y

IV - Annales

Le modèle de régression linéaire simple est défini de la façon suivante avec α et β comme paramètres du modèle :

$$Y = \beta + \alpha X + \varepsilon$$

Dans cette définition

- A. Y est un prédicteur
- B. α s'exprime avec la même unité que celle associée à X
- C. X et Y sont appariées
- D. X et Y correspondent aux inconnues du modèle
- E. ε s'exprime avec la même unité que celle associée à Y

IV - Annales

Le modèle de régression linéaire simple est défini de la façon suivante avec α et β comme paramètres du modèle :

$$Y = \beta + \alpha X + \varepsilon$$

Dans cette définition

- A. Y est un prédicteur.**
→ Le prédicteur est X ! Y est le critère
- B. α s'exprime avec la même unité que celle associée à X**
→ $\text{Unité}(\alpha) = \text{Unité}(Y) / \text{Unité}(X)$
- C. X et Y sont appariées.**
- D. X et Y correspondent aux inconnues du modèle.**
→ C'est alpha et bêta les inconnues.
- E. ε s'exprime avec la même unité que celle associée à Y.**

IV - Annales

Dans un article, un coefficient de corrélation linéaire a été calculé afin de prédire la valeur de la tension artérielle systolique en fonction de la dose du médicament. Vingt doses successives allant de 0,01 mg à 0,20 mg (avec une augmentation régulière de 0,01 mg d'une dose à la dose immédiatement supérieure) ont été conditionnées. Chaque dose a été dispensée à 30 sujets. L'étude a ainsi concerné 600 sujets chez lesquels la tension artérielle systolique a été mesurée en fin de traitement.

En supposant ses conditions d'applications remplies, la démarche statistique la plus adaptée pour répondre au but de l'étude

- A. est celle utilisée dans l'article
- B. serait d'utiliser des comparaisons de moyennes deux à deux
- C. serait d'utiliser un modèle avec la dose comme prédicteur
- D. serait d'utiliser un modèle avec la tension artérielle systolique comme prédicteur
- E. devra prendre en compte un terme d'erreur aléatoire

IV - Annales

Dans un article, un coefficient de corrélation linéaire a été calculé afin de prédire la valeur de la tension artérielle systolique en fonction de la dose du médicament. Vingt doses successives allant de 0,01 mg à 0,20 mg (avec une augmentation régulière de 0,01 mg d'une dose à la dose immédiatement supérieure) ont été conditionnées. Chaque dose a été dispensée à 30 sujets. L'étude a ainsi concerné 600 sujets chez lesquels la tension artérielle systolique a été mesurée en fin de traitement.

En supposant ses conditions d'applications remplies, la démarche statistique la plus adaptée pour répondre au but de l'étude

- A. est celle utilisée dans l'article
- B. serait d'utiliser des comparaisons de moyennes deux à deux
- C. serait d'utiliser un modèle avec la dose comme prédicteur
- D. serait d'utiliser un modèle avec la tension artérielle systolique comme prédicteur
- E. devra prendre en compte un terme d'erreur aléatoire

IV - Annales

A. Est celle utilisée dans l'article.

→ X et Y ne sont pas interchangeables, c'est donc une régression linéaire et non une corrélation (comme dit au début de l'article).

B. Serait d'utiliser des comparaisons de moyennes deux à deux.

→ rien à voir.

C. Serait d'utiliser un modèle avec la dose comme prédicteur.

→ et la tension comme critère.

D. Serait d'utiliser un modèle avec la tension artérielle systolique comme prédicteur.

E. Devra prendre en compte un terme d'erreur aléatoire.

→ Epsilon

Sondage de satisfaction ❤️

