Quelques notions de base
En statistique

Dr. Guillaume Herbet
guillaume.herbet@univmontpellier.fr

Vous êtes CHRU MONTPELLIER CENTRE HOSPITALIER RÉGIONAL UNIVERSITAIRE
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale
UM1 Université Montpellier 1
inm
Association pour la Recherche sur le Cancer

Plan

1- concepts de bases en statistiques

1-1 Les termes importants
1-2 Statistiques descriptives versus inférentielles
1-3 Echelles de mesure
1-4 Quelles statistiques utiliser ?

2- Description et représentations des données

2-1 Représentation graphique des données
2-2 Description des données
2-3 Quelques notations importantes
2-4 Indices de tendance centrale
2-5 Mesures de la variabilité

Plan

3- La distribution normale

3-1 La distribution normale
3-2 La distribution normale centrée réduite.

4- Hypothèses

4-1 Distribution d'échantillonnage
4-2 La théorie du test d'hypothèse
4-3 L'hypothèse nulle
4-4 Statistique de test
4-5 Distribution normale et test d'hypothèse

Plan

5- Le test d'hypothèse appliqué aux moyennes

5-1 Distribution d'échantillonnage de la moyenne et théorème central limite
5-2 Le test t sur un échantillon (σ inconnu)
5-3 Test *t* pour échantillons appariés
5-4 Test *t* pour échantillons indépendants
5-5 Précautions d'application

6- Corrélation et régression linéaire

5-1 Le diagramme de dispersion
5-2 Le concept de covariance
6-3 Le coefficient de corrélation de Pearson
6-4 r ajusté
6-5 Droite de régression
6-6 la signification de r^2
6-7 Test de signification de r et b
6-8 Les facteurs influençant la corrélation

1- Concepts de base en statistiques

1-1 Les termes importants

1-2 Statistiques descriptives versus inférentielles

1-3 Les échelles de mesure

1-4 Quelles statistiques utiliser ?

1- Concepts de base en statistiques

1-1 Les termes importants à retenir

❖ <u>Echantillon</u>

Lorsque vous voulez un phénomène (ex: la dénomination), vous ne pouvez pas étudier ledit phénomène chez tous les sujets qui potentiellement vous intéressent. Par exemple, vous ne pouvez pas étudier la dénomination chez tous les sujets atteints d'une maladie d'Alzheimer, car cela nécessiterait des milliers d'évaluation. Vous devez donc prélever un échantillon d'observations dans la **population parente**.

❖ <u>Echantillon aléatoire</u>

Cette échantillon doit être tiré au hasard pour que celui soit représentatif. Autrement dit, chaque patient atteint de la MA dans notre exemple précédent doit avoir la même probabilité d'appartenir à l'échantillon. Cette échantillon est dit **aléatoire**.

❖ <u>Population parente</u>

La population est l'ensemble des évènements qui vous intéressent (par exemple, le score en dénomination de tous les patients atteints d'une MA).

1- Concepts de base en statistiques

❖ Validité externe

Dans bien des situations, il n’est pas possible d’obtenir un échantillon aléatoire, si bien que les estimations que nous faisons sur la population peuvent ou non refléter avec précision l’ensemble de la population. Ces estimations peuvent donc manquer de **validité externe**.

Exemple : *Vous voulez étudier les processus de lecture chez les patients atteints d’une lésion de l’aire de la forme visuelle des mots. Vous évaluez 5 patients à cet effet. Il est fort probable que votre échantillon manque de validité externe.*

❖ Variables discrètes versus continues

- Les *variables discrètes* sont des variables qui ne prennent qu'un nombre limité de valeurs. Par exemple, le sexe avec deux modalités (homme versus femme).

- Les *variables continues* prennent en principe n’importe quelles valeurs au sein d’une échelle prédéterminée. C’est par exemple le cas de l’âge.

1- Concepts de base en statistiques

❖ Données de mesure versus données quantitatives

La distinction entre variables continues et variables discrètes est intimement liée à celle entre données de mesure (données quantitatives) et données catégorielles (données qualitatives).

- *Données de mesure* : Fait référence aux résultats de tout type de mesure, par exemple le score obtenu à un test, le poids des personnes, etc.

- *Données catégorielles* : Fait référence aux énoncés de type « *dans notre étude, 10 sujets ont obtenu des performances normales et 12 sujets ont obtenu des performances déficitaires* » ; On classe les sujets par catégories.

1- Concepts de base en statistiques

IMPORTANT

❖ Variable indépendante versus variable dépendante

Les variables *indépendantes* sont les variables **manipulées** par l'expérimentateur alors que les *variables dépendantes*, pour faire simple, font **références aux données**.

Exemple : *vous souhaitez comparer les scores en dénomination entre un groupe de patients atteints d'une lésion de l'hémisphère gauche versus des participants qui sont neurologiquement normaux.*

=> *Quel est la variable dépendante ? Quelle est la variable indépendante ?*

Remarque: Les VID peuvent être quantitatives ou qualitatives alors qu'en général (ce n'est pas toujours vrai), les VD sont quantitatives.

STATISTICA - [Data: Spreadsheet2* (3v by 20c)]

	1 Identification Sujets	2 Groupe	3 Score de lecture (/20)
1	SN1	Sujets normaux	20
2	SN2	Sujets normaux	19
3	SN3	Sujets normaux	18
4	SN4	Sujets normaux	16
5	SN5	Sujets normaux	18
6	SN6	Sujets normaux	19
7	SN7	Sujets normaux	20
8	SN8	Sujets normaux	14
9	SN9	Sujets normaux	20
10	SN10	Sujets normaux	17
11	P1	Patients	18
12	P2	Patients	13
13	P3	Patients	12
14	P4	Patients	15
15	P5	Patients	9
16	P6	Patients	12
17	P7	Patients	5
18	P8	Patients	8
19	P9	Patients	12
20	P10	Patients	12

Quel est la variable dépendante ? Quelle est la variable indépendante ?

STATISTICA - [Data: Spreadsheet2* (5v by 20c)]

	1 Identification Sujets	2 Groupe	3 Sexe	4 Age	5 Score de lecture (/20)
1	SN1	Sujets normaux	F	25	20
2	SN2	Sujets normaux	M	32	19
3	SN3	Sujets normaux	M	45	18
4	SN4	Sujets normaux	M	36	16
5	SN5	Sujets normaux	F	18	18
6	SN6	Sujets normaux	F	65	19
7	SN7	Sujets normaux	F	53	20
8	SN8	Sujets normaux	M	30	14
9	SN9	Sujets normaux	F	35	20
10	SN10	Sujets normaux	M	39	17
11	P1	Patients	M	40	18
12	P2	Patients	M	36	13
13	P3	Patients	M	29	12
14	P4	Patients	F	50	15
15	P5	Patients	F	66	9
16	P6	Patients	M	22	12
17	P7	Patients	M	35	5
18	P8	Patients	F	45	8
19	P9	Patients	M	25	12
20	P10	Patients	f	30	12

Quel est la variable dépendante ? Quelles sont les variables indépendantes ?

STATISTICA - [Data: Spreadsheet2* (7v by 20c)]

	1 Identification Sujets	2 Groupe	3 Sexe	4 Age	5 Score de lecture (/20) Avant l'opération	6 Score de lecture (/20) Juste après opération	7 Score de lecture (/20) Trois mois après l'opération
1	SN1	Sujets normaux	F	25	20	18	20
2	SN2	Sujets normaux	M	32	19	20	20
3	SN3	Sujets normaux	M	45	18	18	18
4	SN4	Sujets normaux	M	36	16	17	17
5	SN5	Sujets normaux	F	18	18	18	18
6	SN6	Sujets normaux	F	65	19	19	20
7	SN7	Sujets normaux	F	53	20	19	20
8	SN8	Sujets normaux	M	30	14	15	17
9	SN9	Sujets normaux	F	35	20	19	20
10	SN10	Sujets normaux	M	39	17	18	18
11	P1	Patients	M	40	18	12	14
12	P2	Patients	M	36	18	10	16
13	P3	Patients	M	29	17	10	16
14	P4	Patients	F	50	15	15	17
15	P5	Patients	F	66	20	15	16
16	P6	Patients	M	22	12	12	12
17	P7	Patients	M	35	10	5	11
18	P8	Patients	F	45	15	7	8
19	P9	Patients	M	25	12	9	12
20	P10	Patients	F	30	12	8	10

For Help, press F1 | Spreadsheet2 | C8,V7 | 17 | Sel:OFF | Weight:OFF | CAP | NUM | REC

Quel est la variable dépendante ? Quelles sont les variables indépendantes ?

1- Concepts de base en statistiques

1-1 Les termes importants

1-2 Statistiques descriptives versus inférentielles

1-3 Les échelles de mesure

1-4 Quelles statistiques utiliser ?

1- Concepts de base en statistiques

$\Rightarrow$ *2 concepts importants*

❖ Statistique descriptive

Lorsque nous souhaitons décrire les données, nous avons recours à des statistiques descriptives. Cela permet d'appréhender les données à un niveau global, avant d'entamer toutes procédures de traitement plus élaborées. C'est **une analyse exploratoire**.

- Faire des histogrammes pour représenter les données ;
- Calculer la moyenne, la médiane et le mode ;
- Calculer les indices de dispersion comme les écarts types ;

❖ Statistique inférentielle

Ensemble des méthodes statistiques qui permettent de faire des inférences, c'est à dire vérifier des hypothèses. Par exemple, je teste l'hypothèse que les patients atteints d'une maladie neurodégénérative ont des difficultés de dénomination par rapport à des sujets contrôles.

1- Concepts de base en statistiques

1-1 Les termes importants

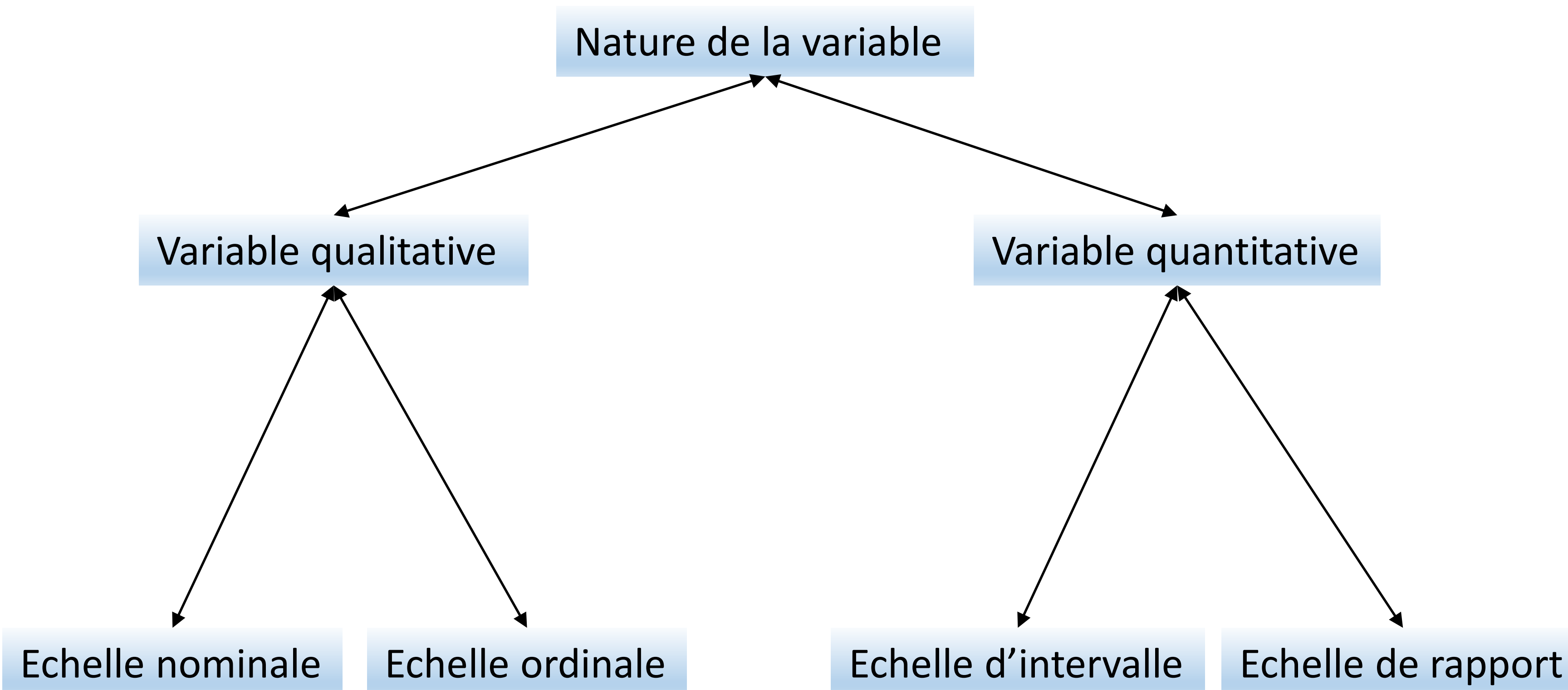
1-2 Statistiques descriptives versus inférentielles

1-3 Les échelles de mesure

1-4 Quelles statistiques utiliser ?

1- Concepts de base en statistiques

⇒ Il existe 4 échelles de mesure différentes



1-3 Les échelles de mesure

1- Concepts de base en statistiques

⇒ Il existe 4 échelles de mesure différentes

❖ <u>Echelle nominale</u>

Les échelles nominales ne sont pas en quelque sorte de vraies échelles dans le sens où il n'y a pas de de relation de hiérarchie entre les dimensions de l'échelle. Les variables comme le sexe sont des variables nominales. On les classe toujours en deux catégories. Ce sont des étiquettes. (autre exemple, la latéralité manuelle).

❖ <u>Echelle ordinale</u>

Les échelles ordinales classent les personnes ou observations selon un continuum prédéfini. Les grades militaires en sont un exemple (Sous-officier< officier < lieutenant <c colonel)

Exemple: Vous faites passer des questionnaires de métacognition à un groupe de sujets. Chaque sujet doit répondre aux affirmations selon l'échelle suivante:

Fort désaccord	Désaccord	D'accord	parfaitement d'accord

1- Concepts de base en statistiques

⇒ Il existe 4 échelles de mesure différentes

- Echelle d’intervalles

Les échelles d’intervalles sont des échelles de mesure pour les quelles on peut à juste titre parler de différences entre les points de l’échelle. L’échelle de Celsius en constitue un bon exemple puisqu’une différence de 5 points a la même signification sur toute l’échelle.

Toutefois, on ne peut pas faire de rapport dans cette échelle (dire que 10° c’est deux fois plus chaud que 5°) dans la mesure ou le 0 est défini de façon arbitraire.

Exemple:

*La pointure = (taille du pieds en CM + 1) * 1,5*

Si taille du pied est de 25 cm, il s’ensuit que la pointure est égal à 39

Pour autant, on ne peut pas dire que quelqu’un qui fait du 42 (27 cm) a un pied deux fois plus grand qu’une personne faisant du 21 (13 cm).

1- Concepts de base en statistiques

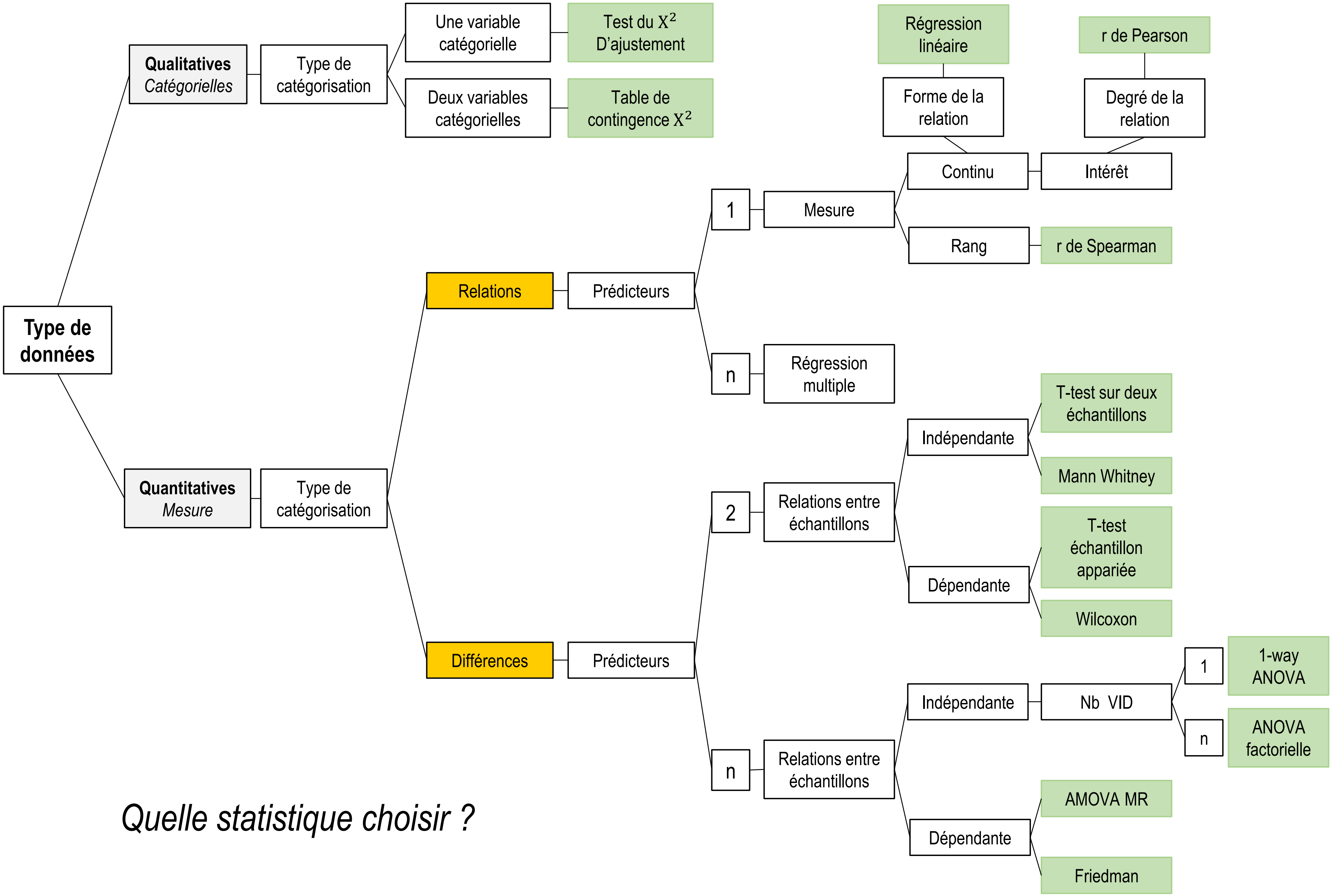
❖ Echelle de rapport

Les échelles de rapport sont des échelles dans lesquelles le point 0 n'est pas défini de façon arbitraire (0 correspond a l'absence d'une chose mesurée)

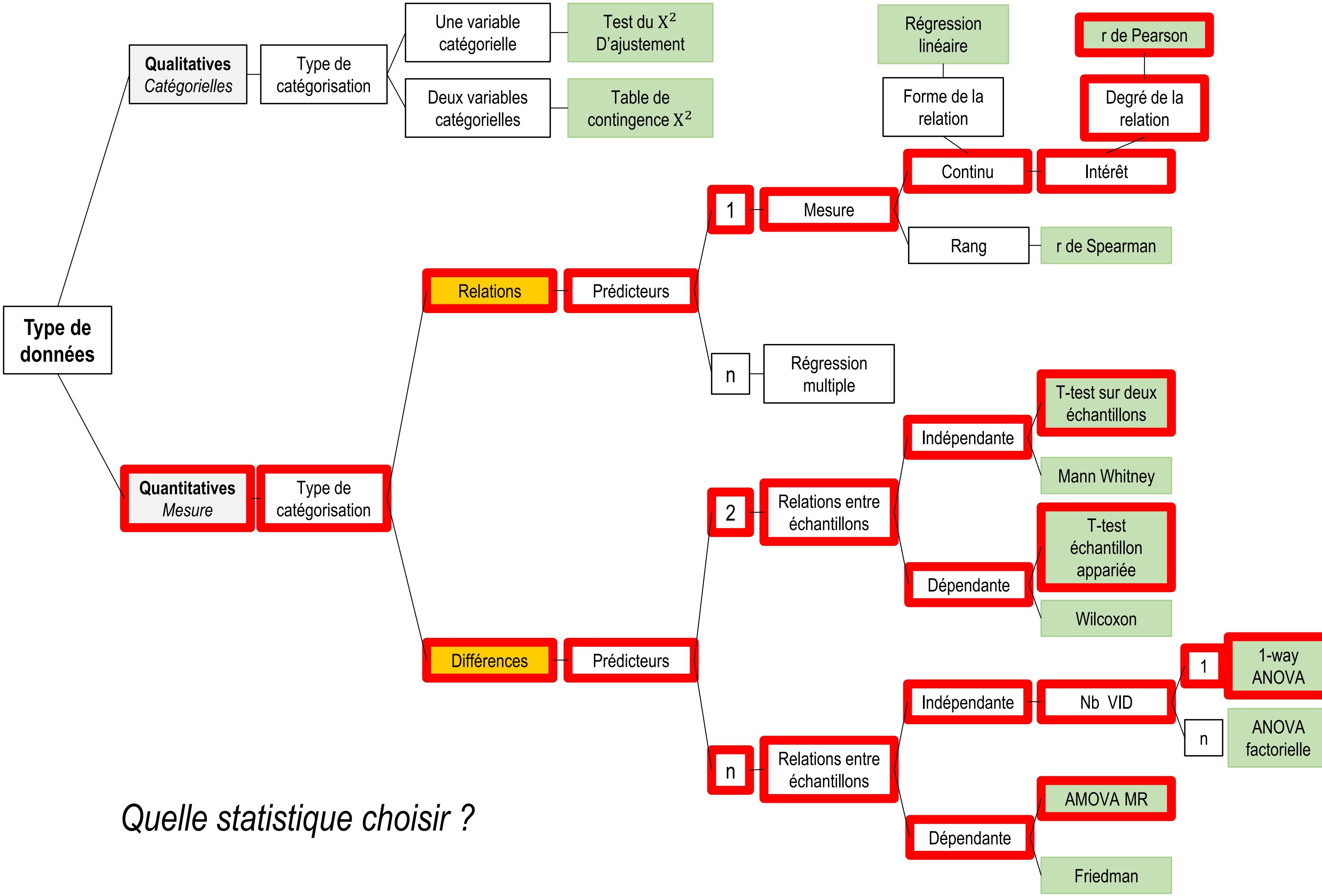
Classiquement, ce sont des échelles physiques habituelles de grandeur telles que la longueur, le poids, les temps de réaction.

1- Concepts de base en statistiques

1-1 Les termes importants
1-2 Statistiques descriptives versus inférentielles
1-3 Les échelles de mesure
1-4 Quelles statistiques utiliser ?



Quelle statistique choisir ?



Quelle statistique choisir ?

2- Distribution et représentation des données

2-1 Représentations graphiques

2-2 Description des données

2-3 Quelques notions importantes

2-4 Indices de tendance centrale

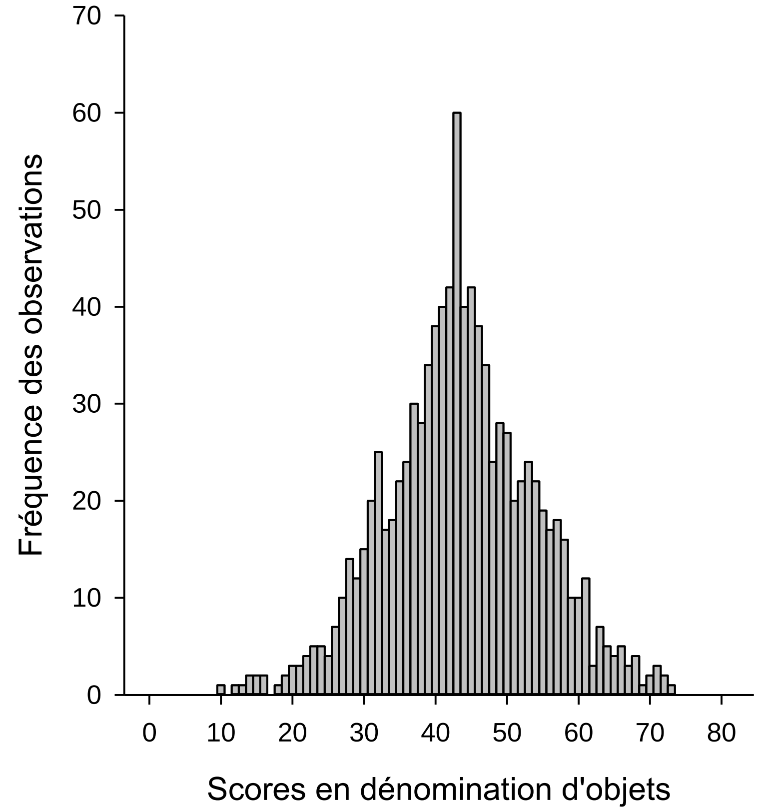
2-5 Mesures de variabilité

2- Description et exploration des données

2-1 Représentation graphique des données

DO Score	Fréquence	DO Score	Fréquence
1	0	41	40
2	0	42	42
3	0	43	60
4	0	44	40
5	0	45	42
6	0	46	38
7	0	47	34
8	0	48	24
9	0	49	28
10	1	50	27
11	0	51	20
12	1	52	22
13	1	53	24
14	2	54	22
15	2	55	19
16	2	56	17
17	0	57	18
18	1	58	16
19	2	59	10
20	3	60	10
21	3	61	12
22	4	62	3
23	5	63	7
24	5	64	5
25	4	65	4
26	7	66	5
27	10	67	3
28	14	68	4
29	12	69	1
30	15	70	2
31	20	71	3
32	25	72	2
33	17	73	1
34	18	74	0
35	22	75	0
36	24	76	0
37	30	77	0
38	28	78	0
39	34	79	0
40	38	80	0

Graphique des scores en dénomination par rapport à la fréquence



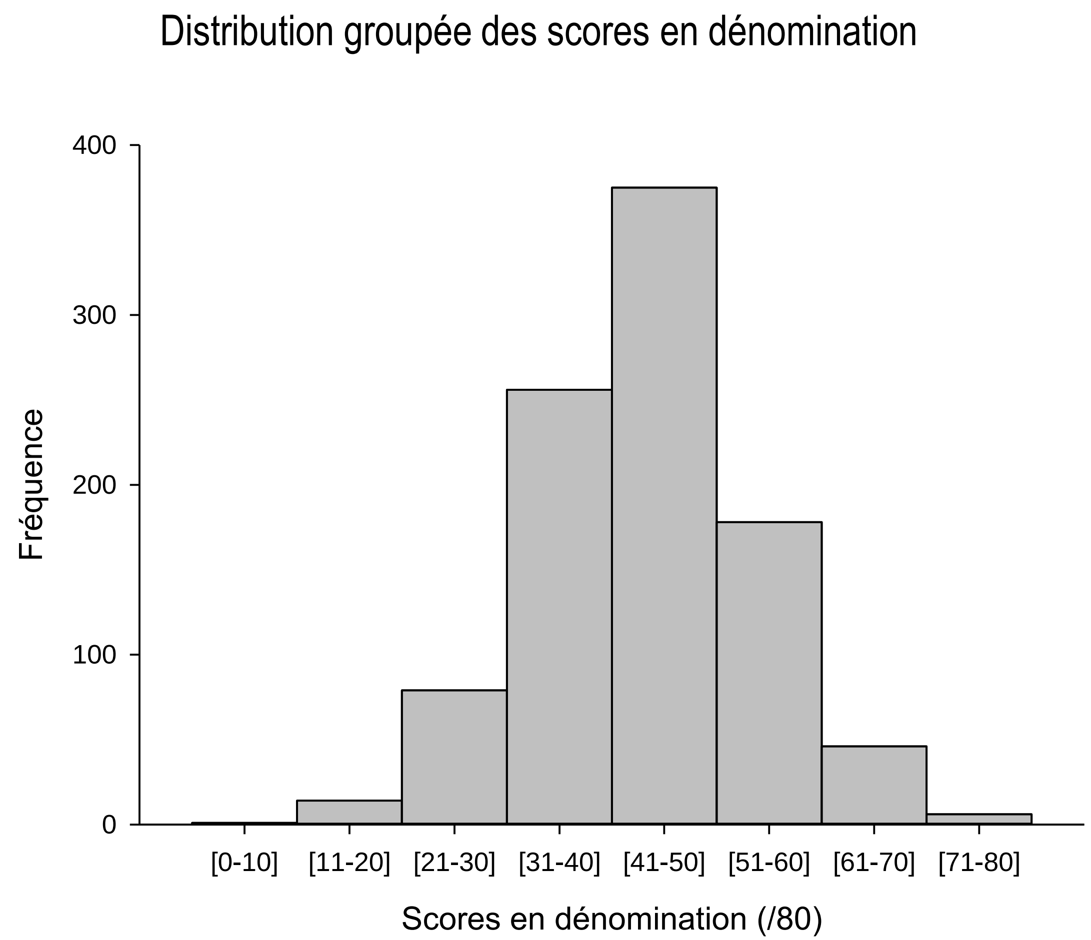
Performances en dénomination de 955 sujets normaux

Histogramme de fréquence des observations

2- Description et exploration des données

Intervalle	Fréquence
[0-10]	1
[11-20]	14
[21-30]	79
[31-40]	256
[41-50]	375
[51-60]	178
[61-70]	46
[71-80]	6

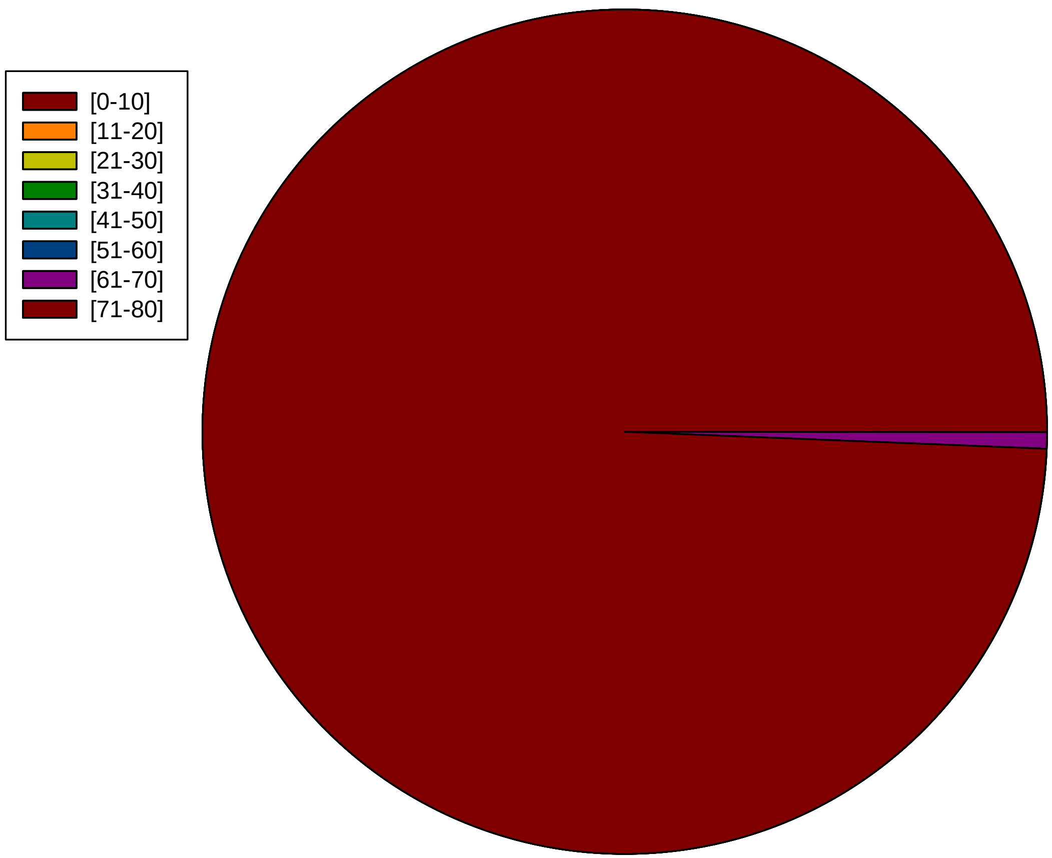
Données triées en intervalles de score
(*même données que la slide précédente*)



Histogramme groupé

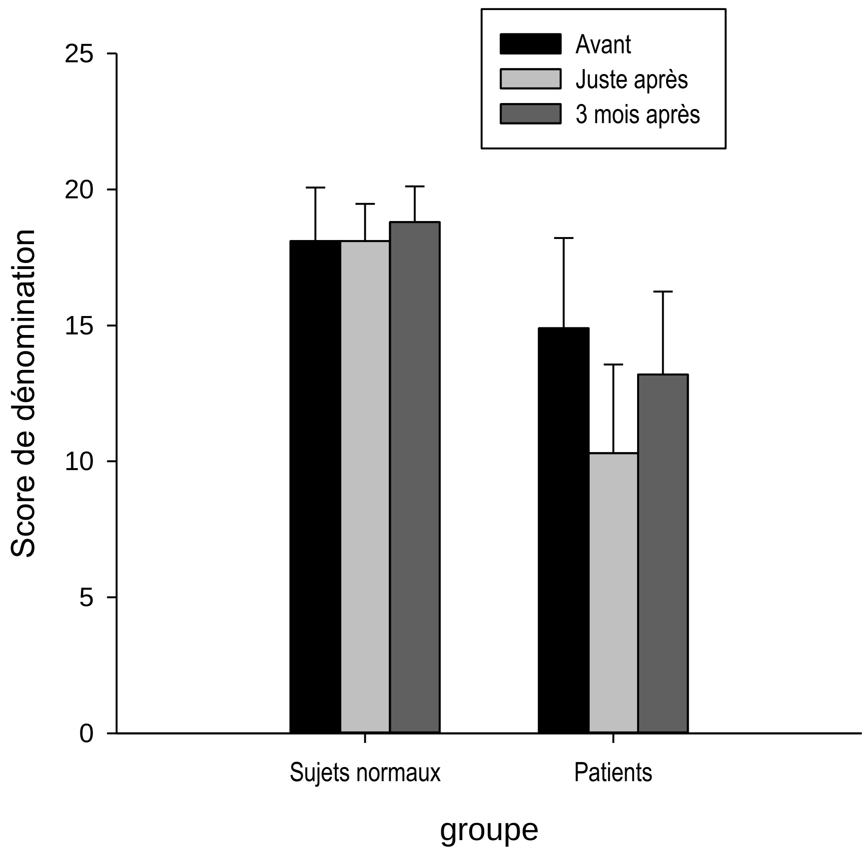
2- Description et exploration des données

2-1 Représentation graphique des données



Camembert des fréquences groupées d'observations

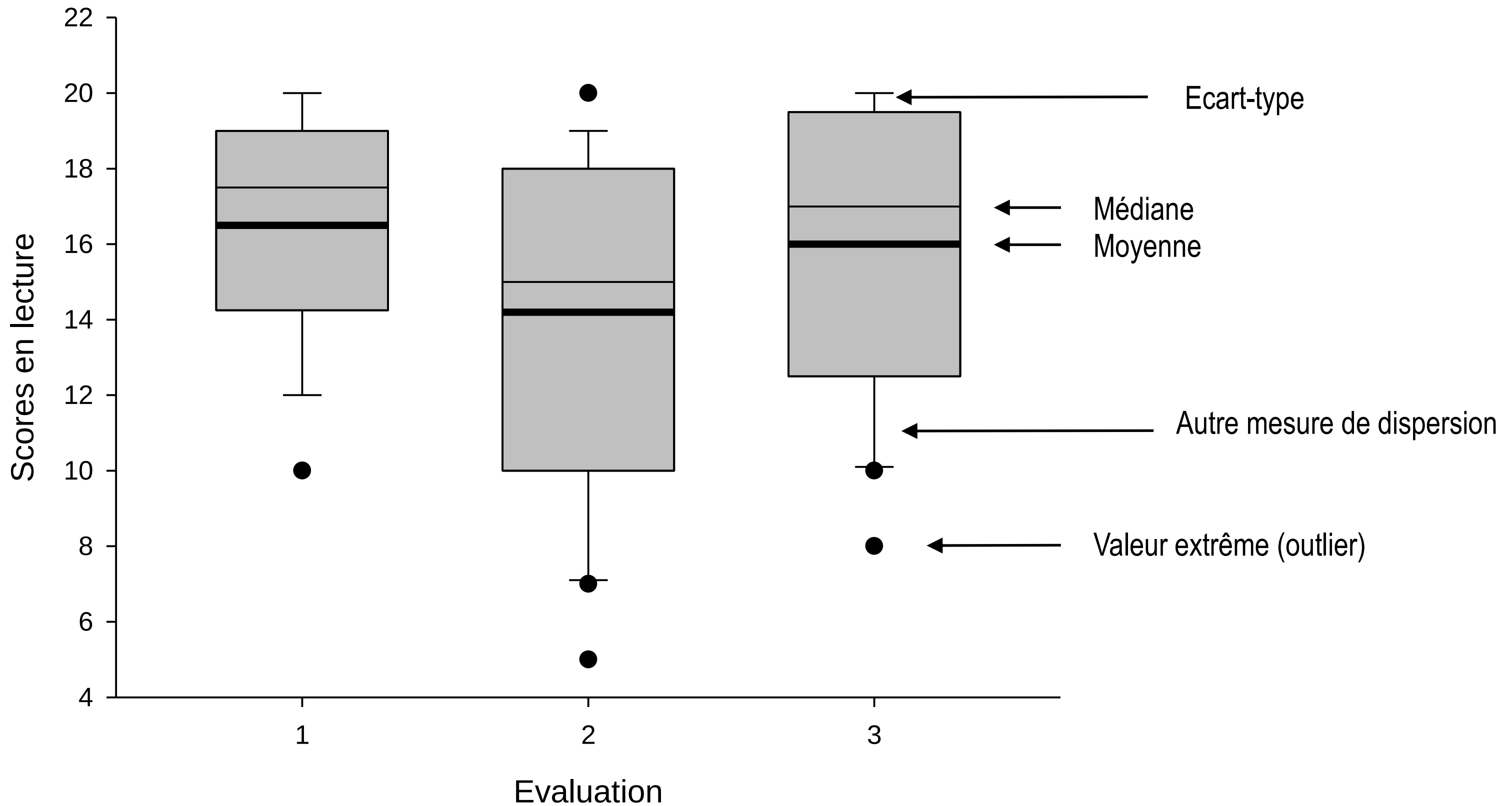
2- Description et exploration des données



Moyennes et écarts-type en fonction du groupe et de l'évaluation

2- Description et exploration des données

2-1 Représentation graphique des données



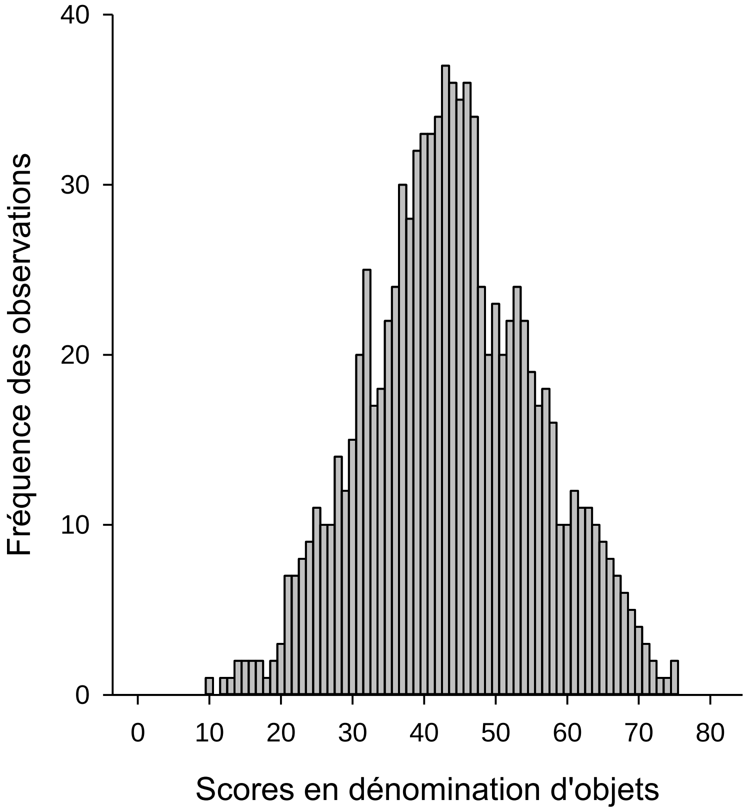
Boite à moustache (ou boxplot) des scores en fonction de l'évaluation

2- Distribution et représentation des données

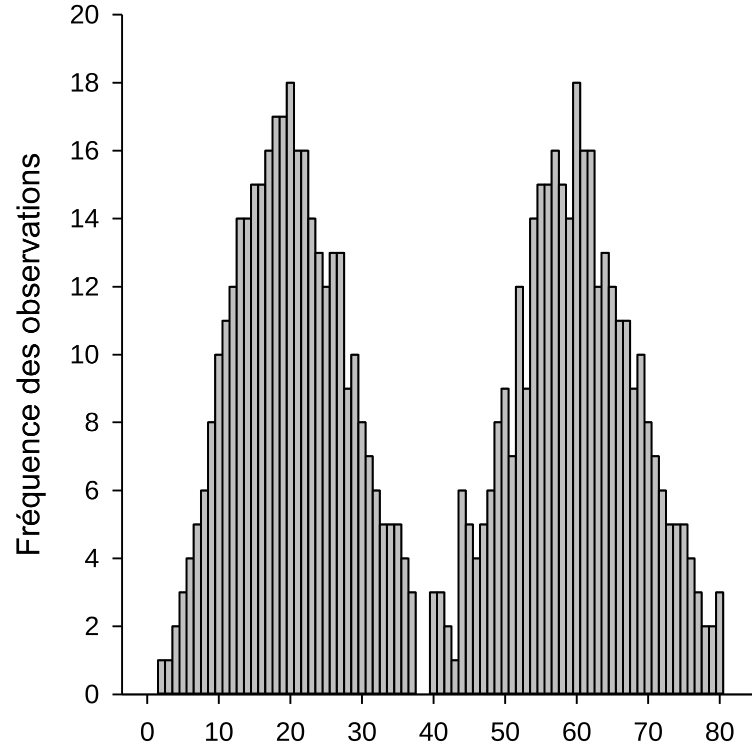
2-1 Représentations graphiques
2-2 Description des données
2-3 Quelques notions importantes
2-4 Indices de tendance centrale
2-5 Mesures de variabilité

2- Description et exploration des données

2-2 Description des distributions de données

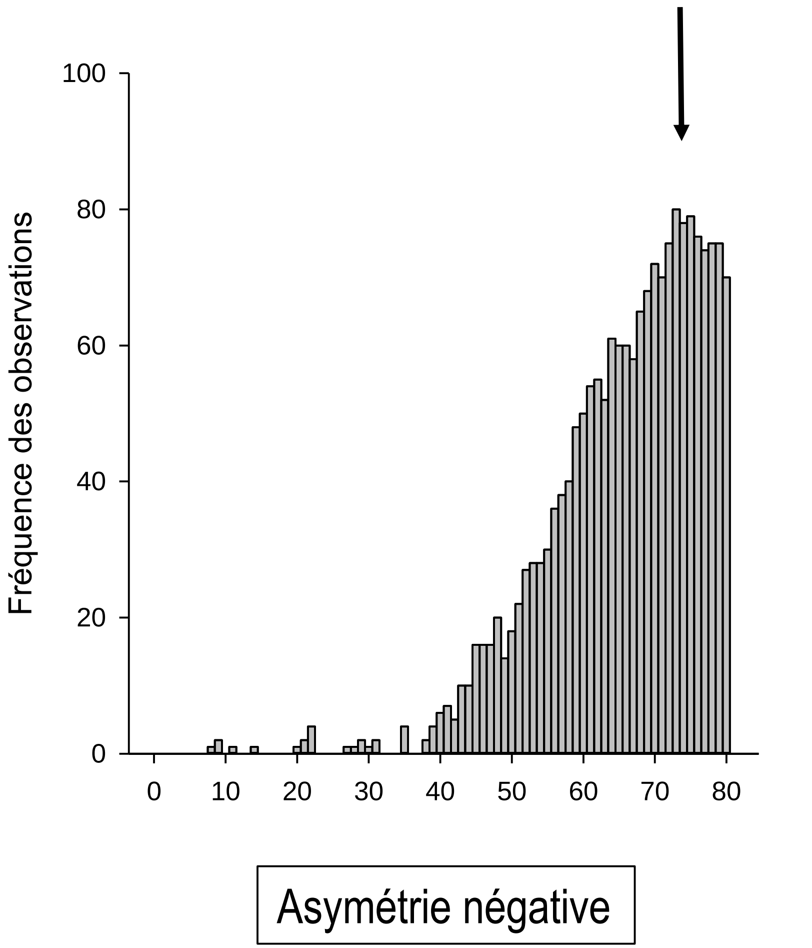


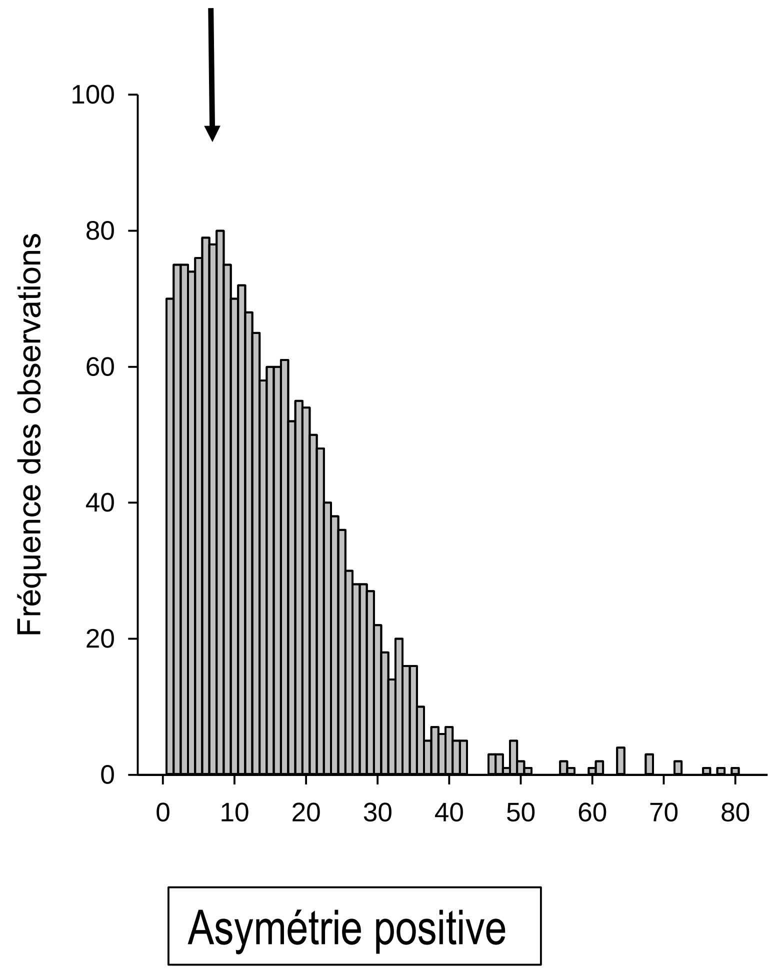
Courbe normale
Ou mésocurtique



Distribution bimodale

2- Description et exploration des données

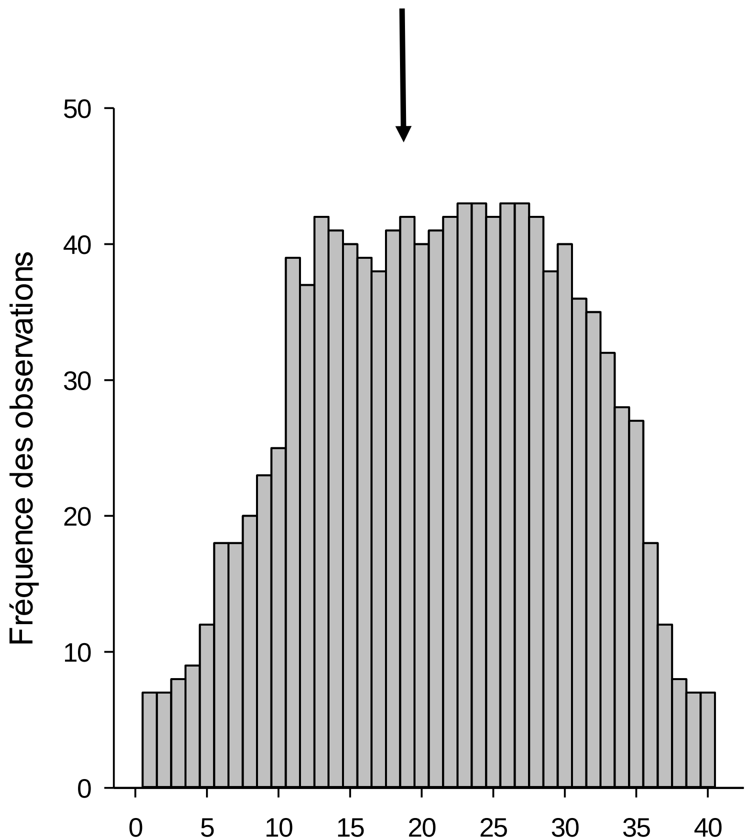




Remarque : il existe un indice pour mesurer l'asymétrie de la courbe : le Skewness

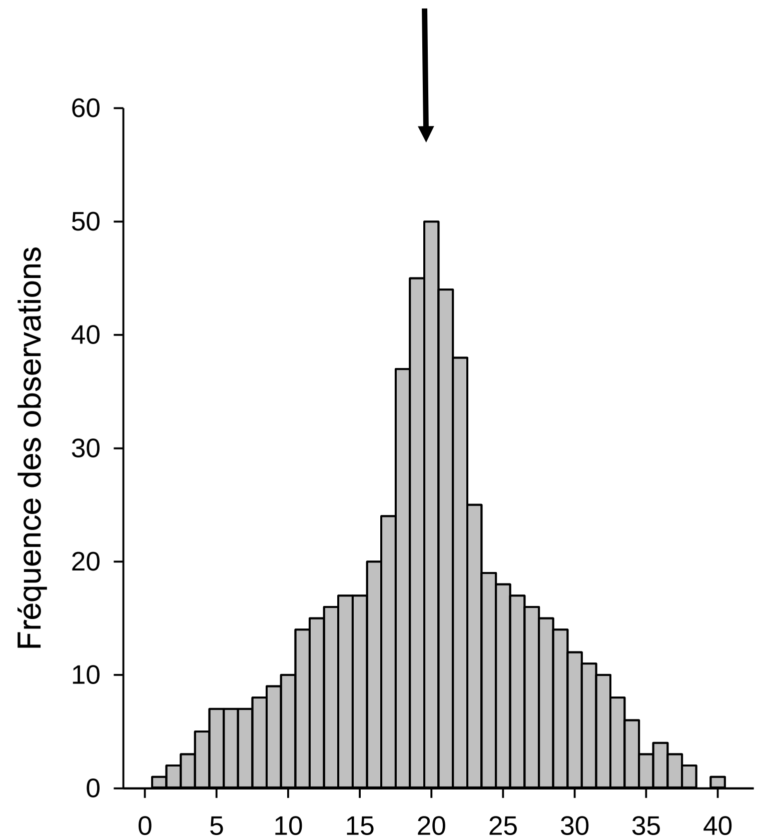
2- Description et exploration des données

La voussure de la courbe est trop plate



Courbe platycurtique

La voussure de la courbe est trop en cloche



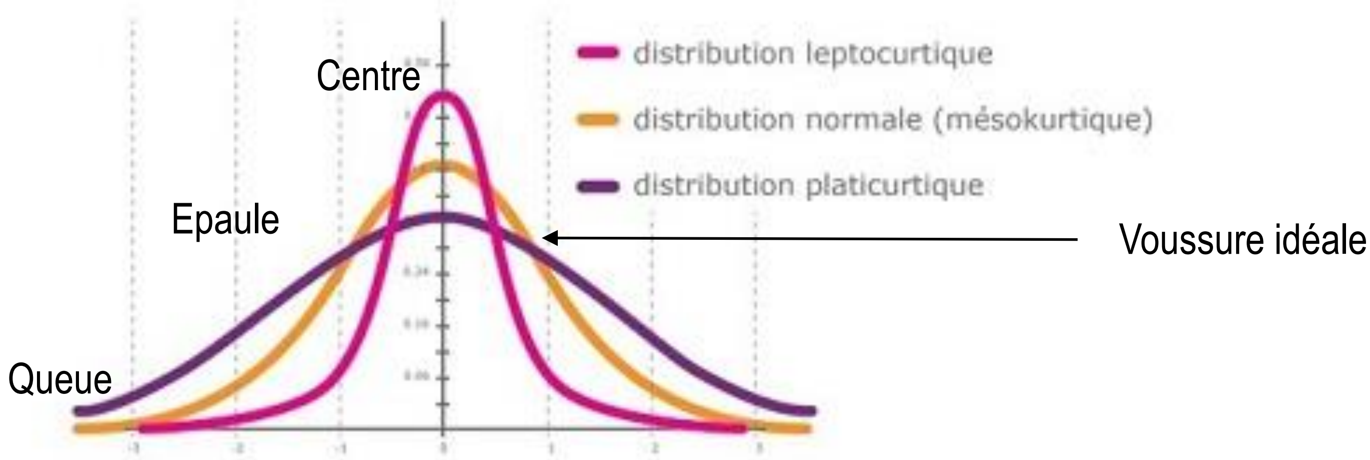
Courbe leptocurtique

Remarque : il existe un indice pour mesurer l’aplatissement de la courbe : le Kurtosis

2- Description et exploration des données

❖ <u>Voussure d'une distribution</u>

La voussure renvoie d'un point de vue mathématique à la concentration relative des scores au centre, dans les extrémités supérieure et inférieure (les queues) et au niveau des épaules entre les centres et les queues.



2- Distribution et représentation des données

2-1 Représentation graphiques

2-2 Description des données

2-3 Quelques notations importantes

2-4 Indices de tendance centrale

2-5 Mesures de variabilité

2- Description et exploration des données

2-3 Quelques notations importantes

❖ Les notations de variables

- Soit une variable (le score de lecture de 6 patients), la notation *X* ou *Y* servira à représenter cette variable. Par exemple, si les scores des patients est [18, 20, 15, 18, 19, 14], X représentera l'ensemble de ces valeurs.

- X_1 fera référence à 18; X_3, fera référence à 15, *etc*.

- X_i peut prendre n'importe quelle valeur entre 1 et 6.

❖ Notation de la somme (Sigma)

$\sum X_i$ « *additionner les X_i* »

$\sum_{i=1}^{N} X_i$ « additionner tous les X_i de $i = 1$ à $i = N$ »

2- Description et exploration des données

Score de lecture (10)	Score de compréhension (5)				
(*X*)	(*Y*)	(X^2)	(Y^2)	***X-Y***	***XY***
8	3	64	9	5	24
6	5	36	25	1	30
10	2	100	4	8	20
9	4	81	81	5	36
5	4	25	16	1	20

$\sum X_i = (8+6+10+9+5) = 38$

$\sum Y_i = (3+5+2+4+4) = 18$

$\sum X_i^2 = (64+36+100+81+25) = 206$

$\sum Y_i^2 = (9+25+4+81+16) = 125$

$\sum(Xi - Yi) = (5+1+8+5+1) = 20$

$\sum X_i Y_i = (24+30+20+36+20) = 130$

$(\sum X_i)^2 = (38)^2 = 1444$

2-3 Quelques notations importantes

2- Distribution et représentation des données

2-1 Représentations graphiques

2-2 Description des données

2-3 Quelques notions importantes

2-4 Indices de tendance centrale

2-5 Mesures de variabilité

2- Description et exploration des données

❖ <u>Double indice</u>

	Score en lecture (10)		
	Evaluation 1	Evaluation 2	***Total***
1	8	7	***15***
2	5	5	***10***
3	10	9	***19***
4	5	6	***11***
5	7	8	***15***
Total	***35***	***35***	

Dans ce type de tableau, *i* va représenter la ligne et *j* la colonne. Si vous voulez faire référence à une case spécifique de tableau, vous allez utiliser la notation X_{ij}

i allant de 1 à 5 et *j* allant de 1 à 2.

$X_{5,2} = 8$

La somme de toutes les données du tableau va s'écrire :

$$\sum_{i=1}^{5}\sum_{j=1}^{2}X_{ij}$$

2- Description et exploration des données

❖ Définition formelle

La mesure de la tendance centrale (ou mesures de position), fait référence à l'ensemble des mesures **liées à l'endroit où la distribution est centrée sur l'échelle** (où se situe le centre de la distribution ?).

Ces mesure sont dépendantes de l'échelle utilisée, et sont sensibles pour certaines d'entre elles, aux valeurs extrêmes de la distribution (*outliers*).

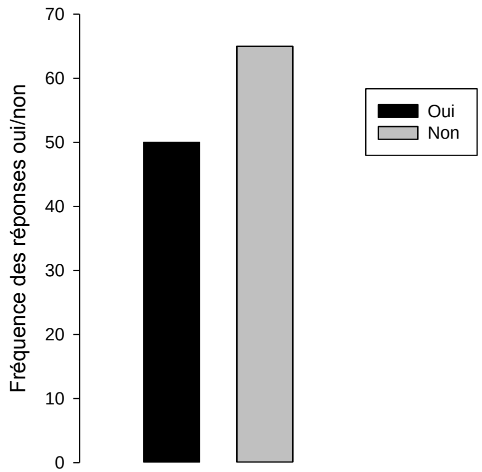
Il existe 3 indices de tendance centrale : le *mode*, la *médiane* et la *moyenne.*

❖ Le mode

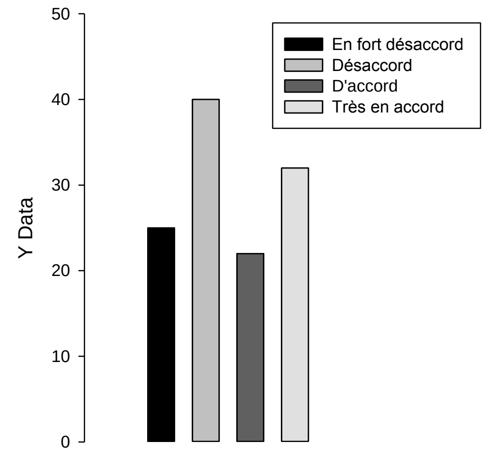
Le mode **est la valeur la plus fréquente** dans une distribution, c'est-à-dire la valeur ou le score pris par le plus grand nombre de participants. Ainsi, le mode est la valeur de X qui correspond au point de plus élevé de la distribution.

Le mode peut-être utilisé dans n'importe quelle échelle, mais est le seul indice possible dans les échelles catégorielles.

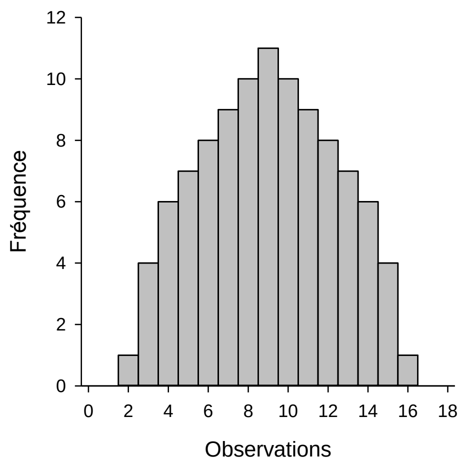
2- Description et exploration des données



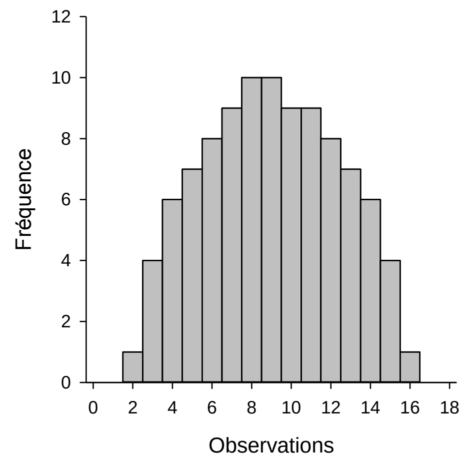
Le mode est « Oui » (65)



Le mode est « Désaccord » (40)



Le mode est 11



Le mode est 8,5

2-4 Mesures de tendance centrale

2- Description et exploration des données

❖ La médiane

D'un point de vue formel, c'est le score qui correspond au point auquel ou sous lequel des scores sont inclus lorsque les données sont disposée en ordre numérique. La position médiane dans une distribution d'écrit donc :

$$\frac{N+1}{2}$$

Patient	N°	Score (10)
Patient 1	1	0
Patient 2	2	2
Patient 3	3	2
Patient 4	4	2
Patient 5	5	3
Patient 6	6	3
Patient 7	7	4
Patient 8	8	4
Patient 9	9	5
Patient 10	**10**	**5**
Patient 11	11	6
Patient 12	12	7
Patient 13	13	7
Patient 14	14	8
Patient 15	15	8
Patient 16	16	8
Patient 17	17	9
Patient 18	18	9
Patient 19	19	10

Position Médiane = (19+1)/2 = 10

Médiane = 5

2- Description et exploration des données

❖ La médiane

D'un point de vue formel, c'est le score qui correspond au point auquel sous lequel des scores sont inclus lorsque les données sont disposée en ordre numérique. La position médiane dans une distribution d'écrit donc :

Patient	N°	Score (10)
Patient 1	1	0
Patient 2	2	2
Patient 3	3	2
Patient 4	4	2
Patient 5	5	3
Patient 6	6	3
Patient 7	7	4
Patient 8	8	4
Patient 9	9	5
Patient 10	10	5
Patient 11	11	6
Patient 12	12	7
Patient 13	13	7
Patient 14	14	8
Patient 15	15	8
Patient 16	16	8
Patient 17	17	9
Patient 18	18	9
Patient 19	19	10
Patient 20	20	10

$$\frac{N+1}{2}$$

Position Médiane = (20+1)/2 = 10,5

Médiane = 5,5

2- Description et exploration des données

❖ La moyenne

La moyenne, indice de tendance centrale le plus utilisé, est la somme des scores divisée par le nombre de scores.

La notation est $\bar{X}$

Et $\bar{X} = \frac{\sum X}{N}$

Si l'on considère cet ensemble de valeur [12, 6, 5, 4, 2], la moyenne est égale à :

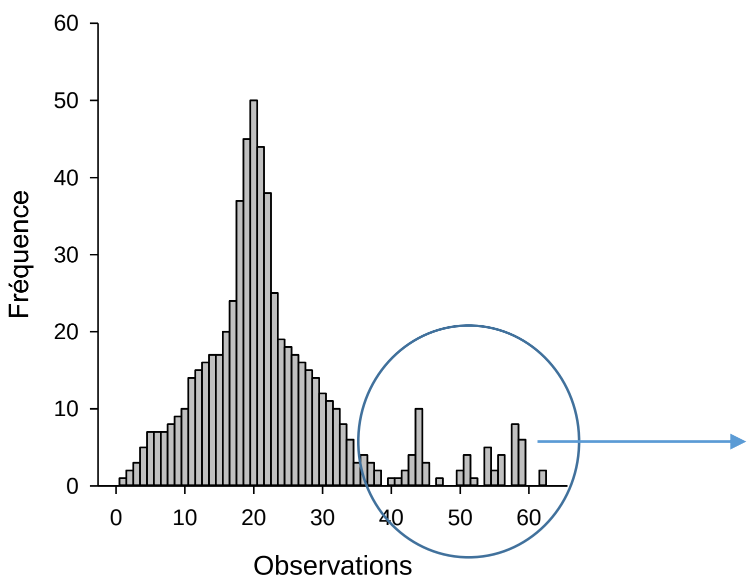
$\frac{12+6+5+4+2}{5}$ = 5.8

2- Description et exploration des données

❖ Quelques remarques

- R1: La moyenne et la médiane ne seront égales que si la distribution est symétrique, et les trois indices de tendance que si la distribution est symétrique et unimodale.

- R2: La médiane n'est pas influencé par les scores extrêmes contrairement à la moyenne.



Données extrêmes qui vont influencées la moyenne (la décaler vers la droite)

2- Description et exploration des données

Patient	temps de réaction (ms)
Patient 1	540
Patient 2	623
Patient 3	750
Patient 4	422
Patient 5	453
Patient 6	472
Patient 7	650
Patient 8	524
Patient 9	583
Patient 10	701
Patient 11	423
Patient 12	499
Patient 13	523
Patient 14	522
Patient 15	605
Patient 16	702
Patient 17	621
Patient 18	530
Patient 19	589

Patient	temps de réaction (ms)
Patient 1	540
Patient 2	623
Patient 3	750
Patient 4	422
Patient 5	453
Patient 6	472
Patient 7	650
Patient 8	524
Patient 9	583
Patient 10	701
Patient 11	423
Patient 12	499
Patient 13	523
Patient 14	522
Patient 15	605
Patient 16	702
Patient 17	621
Patient 18	530
Patient 19	3523

Médiane = 701

Moyenne = 564 ms

Médiane = 701

Moyenne = 719 ms

2- Distribution et représentation des données

2-1 Représentation graphiques

2-2 Description des données

2-3 Quelques notions importantes

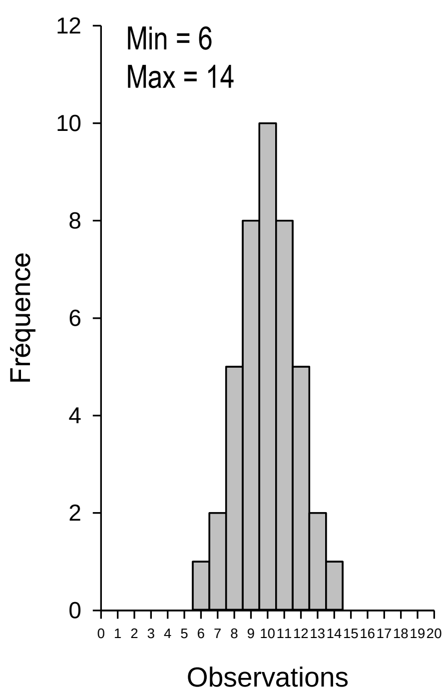
2-4 Indices de tendance centrale

2-5 Mesures de variabilité

2- Description et exploration des données

❖ <u>Dispersion</u>

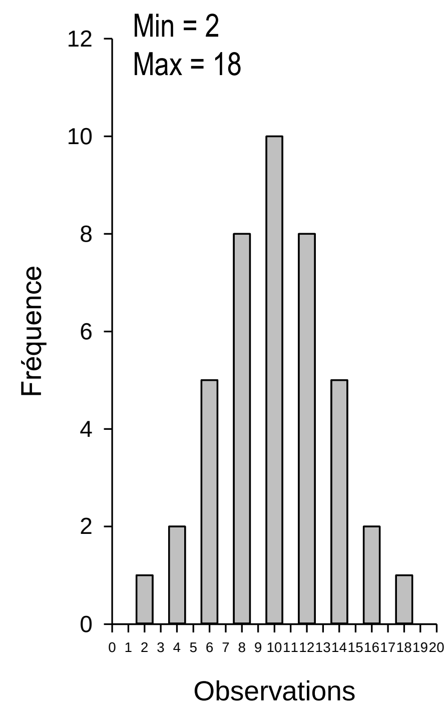
Les indices de tendance centrale permettent de caractériser les données, de les décrire, mais ils ne sont pas suffisamment en en eux-mêmes, pour appréhender, **nous renseigner de façon complète**, la distribution des données.



Mode = 10

Moyenne = 10

Médiane = 10



Pourtant la variabilité autour des indices de tendance est bien plus grande sur la courbe de gauche (deux fois plus grandes)

LA DISPERTION N'EST PAS LA MÊME

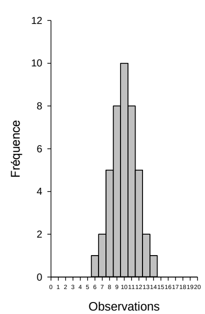
2- Description et exploration des données

❖ <u>Dispersion</u>

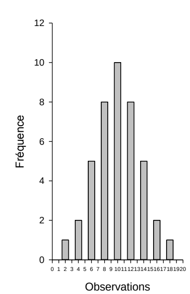
Plusieurs indices de dispersion/variabilité peuvent être utilisées, comme :

- L’étendue
- L’espace interquartile
- L’écart-moyen
- L’écart absolu moyen
- La variance
- L’écart-type

Par exemple: l’étendue est la distance entre la le score la valeur la moins élevée et la valeur la plus élevée de la distribution.



Etendue = 14-6 = 8



Etendue = 18-2 = 16

2- Description et exploration des données

❖ Ecart moyen (écart à la moyenne)

Pour mesurer le degré de dispersion des scores autour de la moyenne (c'est-à-dire leur écart à la moyenne), on pourrait calculer tous les écarts à la moyenne et en faire la moyenne. Mais la moyenne des écarts à la moyenne (leur somme a priori) est toujours égal à 0. Nous ne pouvons utiliser cet écart moyen comme indice de dispersion.

	X	***X- $\bar{X}$***
	5	-5
	6	-4
	7	-3
	8	-2
	9	-1
	10	0
	11	1
	12	2
	13	3
	14	4
	15	5
Moyenne	**10**	**0**

2- Description et exploration des données

❖ La variance de l'échantillon

La variance est un concept clé et une statistique régulièrement utilisée, mais elle n'est pas facile à interpréter. Elle s'écrit :

$$s_X^2 = \frac{\sum(X - \bar{X})^2}{N - 1}$$

	X	$X - \bar{X}$	$(X - \bar{X})^2$
	5	-5	25
	6	-4	20
	7	-3	9
	8	-2	4
	9	-1	1
	10	0	0
	11	1	1
	12	2	4
	13	3	9
	14	4	16
	15	5	25
Moyenne	10	0	10,36

$$s_X^2 = \frac{\sum(X - \bar{X})^2}{N - 1} = \frac{(-5)^2 + \ldots + 5^2}{10 - 1} = \frac{114}{9} = 12,66$$

❖ Ecart-type

L'écart type (s) se définit comme la *racine carrée positive de la variance*. Elle est parfois appelée **déviation standard**. C'est la mesure de dispersion la plus utilisée et la plus facilement interprétable. Formellement :

$$s_X = \sqrt{\frac{\sum(X - \bar{X})^2}{N - 1}}$$

Si nous reprenons l'exemple d'avant, l'écart type est donc de :

$s_X = \sqrt{12{,}66} = 3{,}56$

Remarque : le calcul de la variance et de l'écart-type se base sur les données de l'échantillon. Si vous voulez faire une estimation sur la population, vous retirez le 1 à la formule. s_X^2 devient alors σ_X^2 ; s_X devient σ_X

2- Description et exploration des données

❖ <u>Ecart-type</u>

Pour caractériser une distribution dans un mémoire ou dans un article de recherche, on note généralement de la façon suivante :

Moyenne ± Ecart-type [min-max]

Par exemple,

10 ± 3,56 [5-15]

	X	***X***- $\bar{X}$	***(X***- $\bar{X}$)²
	5	-5	25
	6	-4	20
	7	-3	9
	8	-2	4
	9	-1	1
	10	0	0
	11	1	1
	12	2	4
	13	3	9
	14	4	16
	15	5	25
Moyenne	10	0	10,36

2- Description et exploration des données

❖ Ecart-type

L'écart-type est particulièrement sensible aux valeurs extrêmes, raison pour laquelle il représente un bon indice de dispersion mais seulement quand les données suivent une distribution normale.

Patient	temps de réaction (ms)
Patient 1	540
Patient 2	623
Patient 3	750
Patient 4	422
Patient 5	453
Patient 6	472
Patient 7	650
Patient 8	524
Patient 9	583
Patient 10	701
Patient 11	423
Patient 12	499
Patient 13	523
Patient 14	522
Patient 15	605
Patient 16	702
Patient 17	621
Patient 18	530
Patient 19	589

Médiane = 701 ms

Moyenne = 564 ms

ET = 94 ms

Patient	temps de réaction (ms)
Patient 1	540
Patient 2	623
Patient 3	750
Patient 4	422
Patient 5	453
Patient 6	472
Patient 7	650
Patient 8	524
Patient 9	583
Patient 10	701
Patient 11	423
Patient 12	499
Patient 13	523
Patient 14	522
Patient 15	605
Patient 16	702
Patient 17	621
Patient 18	530
Patient 19	3523

Médiane = 701 ms

Moyenne = 719 ms

ET = 685 ms

Médiane = 701

Moyenne = 564 ms

Moyenne = 719 ms

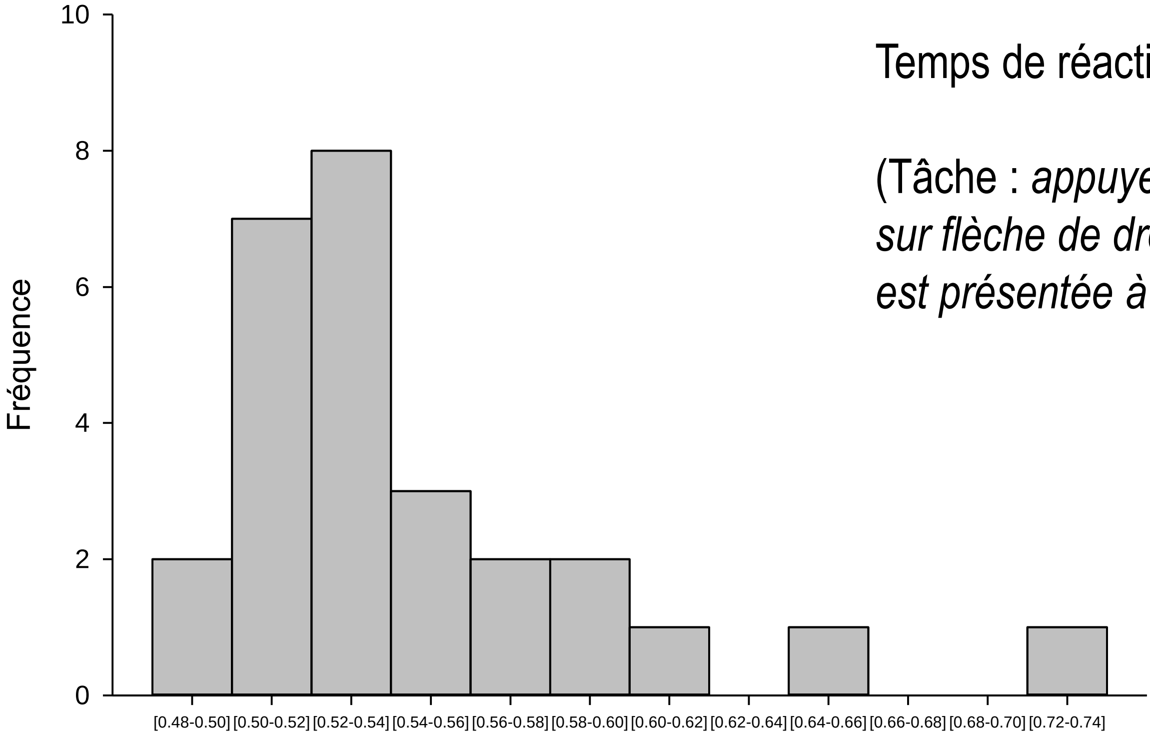
3- La distribution normale

3-1 La distribution normale

3-2 Description normale centrée réduite

3- La distribution normale

❖ La distribution normale est l'une des distributions que l'on rencontre le plus (dans la nature), et ce d'autant plus quand on étudie des processus psychologiques.



Temps de réaction d'un patient

(Tâche : *appuyer le plus rapidement possible sur flèche de droite quand une flèche droite est présentée à l'écran et vice-versa*).

Distribution quasi normale, légèrement asymétrique avec deux outliers

3- La distribution normale

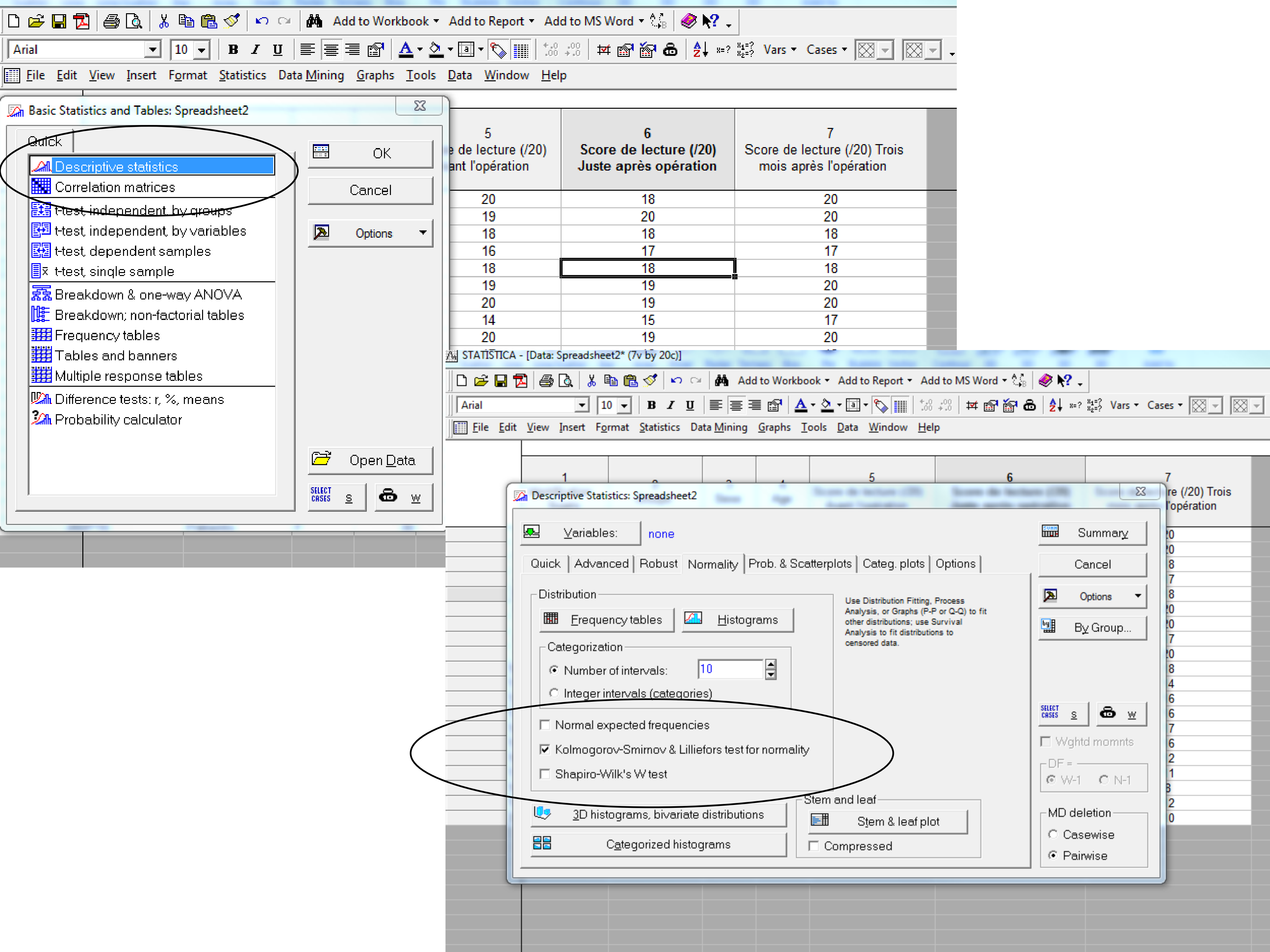
❖ La distribution normale est l'une des distributions que l'on rencontre le plus (dans la nature), et ce d'autant plus quand on étudie des processus psychologiques. Plusieurs raisons à cela :

- Parmi les variables dépendantes qui nous occupent, beaucoup sont supposés être normalement distribuées dans la population. Nous supposons souvent que si nous obtenions l'ensemble des données de la population, celles-ci seraient normales.

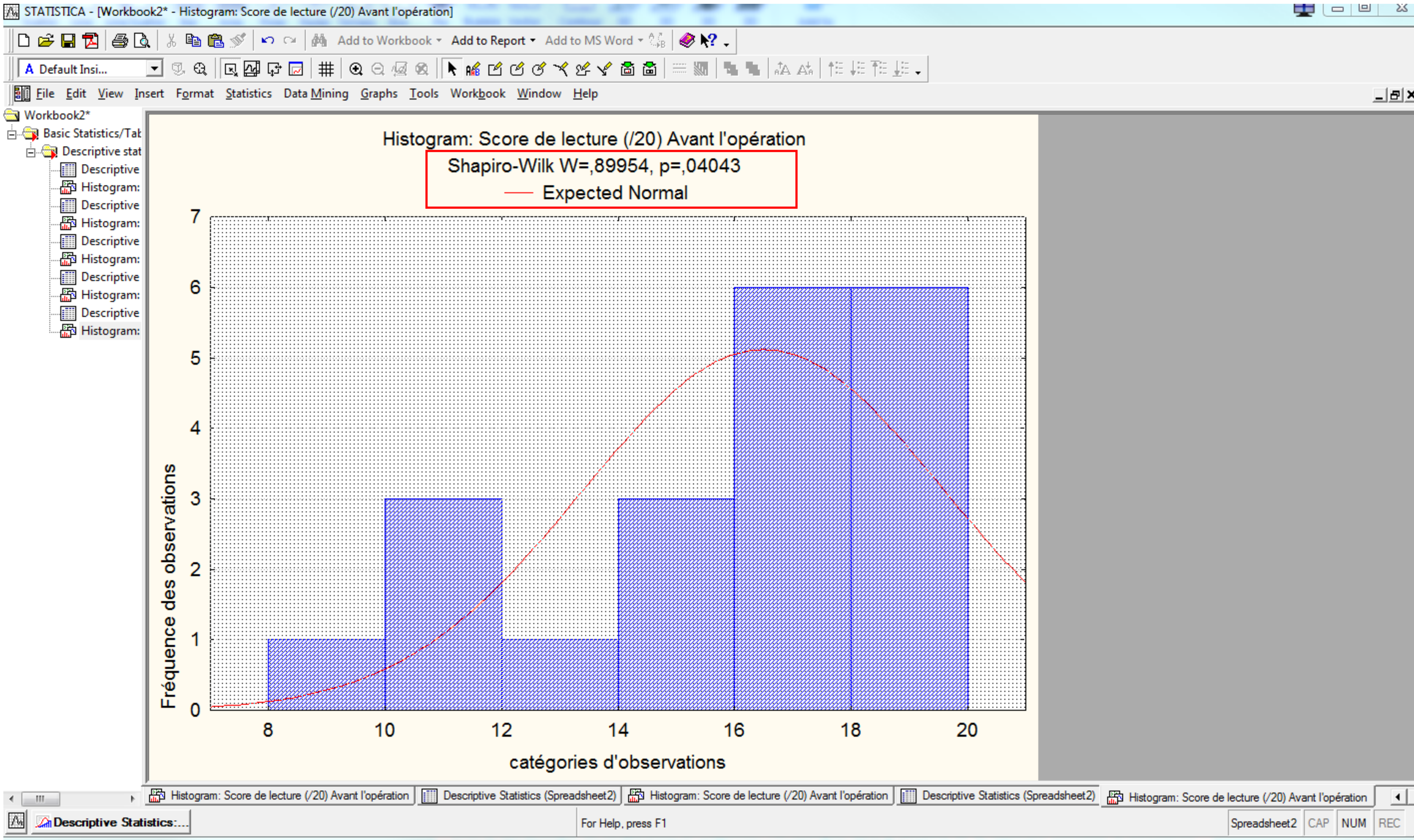
- Si il est supposé que les données sont normales, nous pouvons faire des inférences statistiques plus facilement que si elles n'étaient pas normales (sinon **statistiques non-paramétriques**)

- La plupart des procédures statistiques impliquent d'une manière ou d'une autres que les données soient normales.

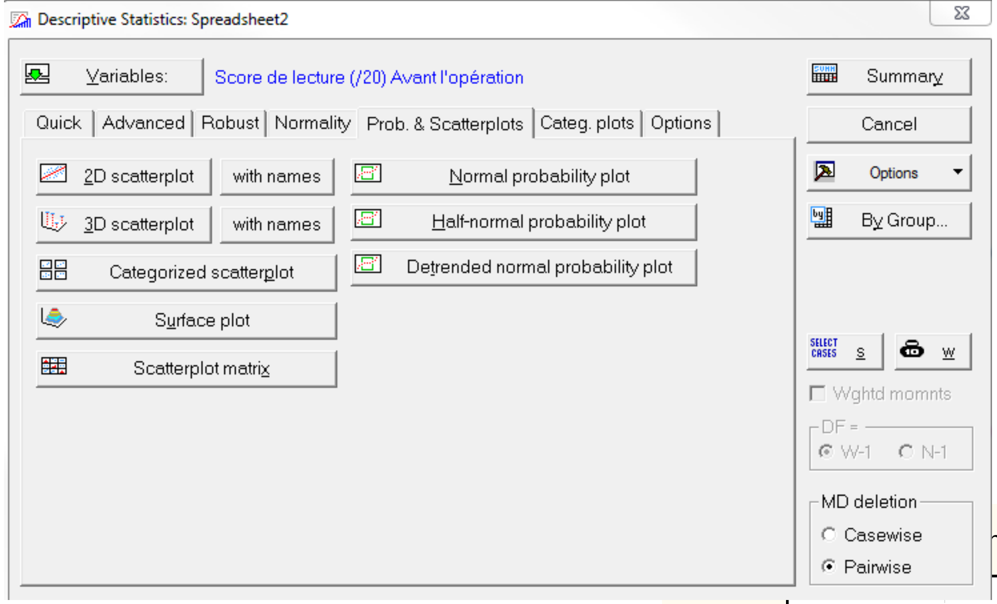
Conséquence: il est obligatoire de procéder avant tout test statistique à quelques vérifications concernant la normalité ou non des données (par exemple, le test de Shapiro-Wilk ou le test de kolmogorov-Smirnov)


Add to Workbook
Add to Report
Add to MS Word
Arial
10
Vars
Cases
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
Basic Statistics and Tables: Spreadsheet2
Quick
Descriptive statistics
Correlation matrices
t-test, independent, by groups
t-test, independent, by variables
t-test, dependent samples
t-test, single sample
Breakdown & one-way ANOVA
Breakdown; non-factorial tables
Frequency tables
Tables and banners
Multiple response tables
Difference tests: r, %, means
Probability calculator
OK
Cancel
Options
Open Data
SELECT CASES
S
W
5
e de lecture (/20)
ant l'opération
6
Score de lecture (/20)
Juste après opération
7
Score de lecture (/20) Trois
mois après l'opération
20 18 20
19 20 20
18 18 18
16 17 17
18 18 18
19 19 20
20 19 20
14 15 17
20 19 20
STATISTICA - [Data: Spreadsheet2* (7v by 20c)]
Descriptive Statistics: Spreadsheet2
Variables:
none
Summary
Cancel
Options
By Group...
Quick
Advanced
Robust
Normality
Prob. & Scatterplots
Categ. plots
Options
Distribution
Frequency tables
Histograms
Use Distribution Fitting, Process Analysis, or Graphs (P-P or Q-Q) to fit other distributions; use Survival Analysis to fit distributions to censored data.
Categorization
Number of intervals:
10
Integer intervals (categories)
Normal expected frequencies
Kolmogorov-Smirnov & Lilliefors test for normality
Shapiro-Wilk's W test
3D histograms, bivariate distributions
Categorized histograms
Stem and leaf
Stem & leaf plot
Compressed
Wghtd momnts
DF =
W-1
N-1
MD deletion
Casewise
Pairwise

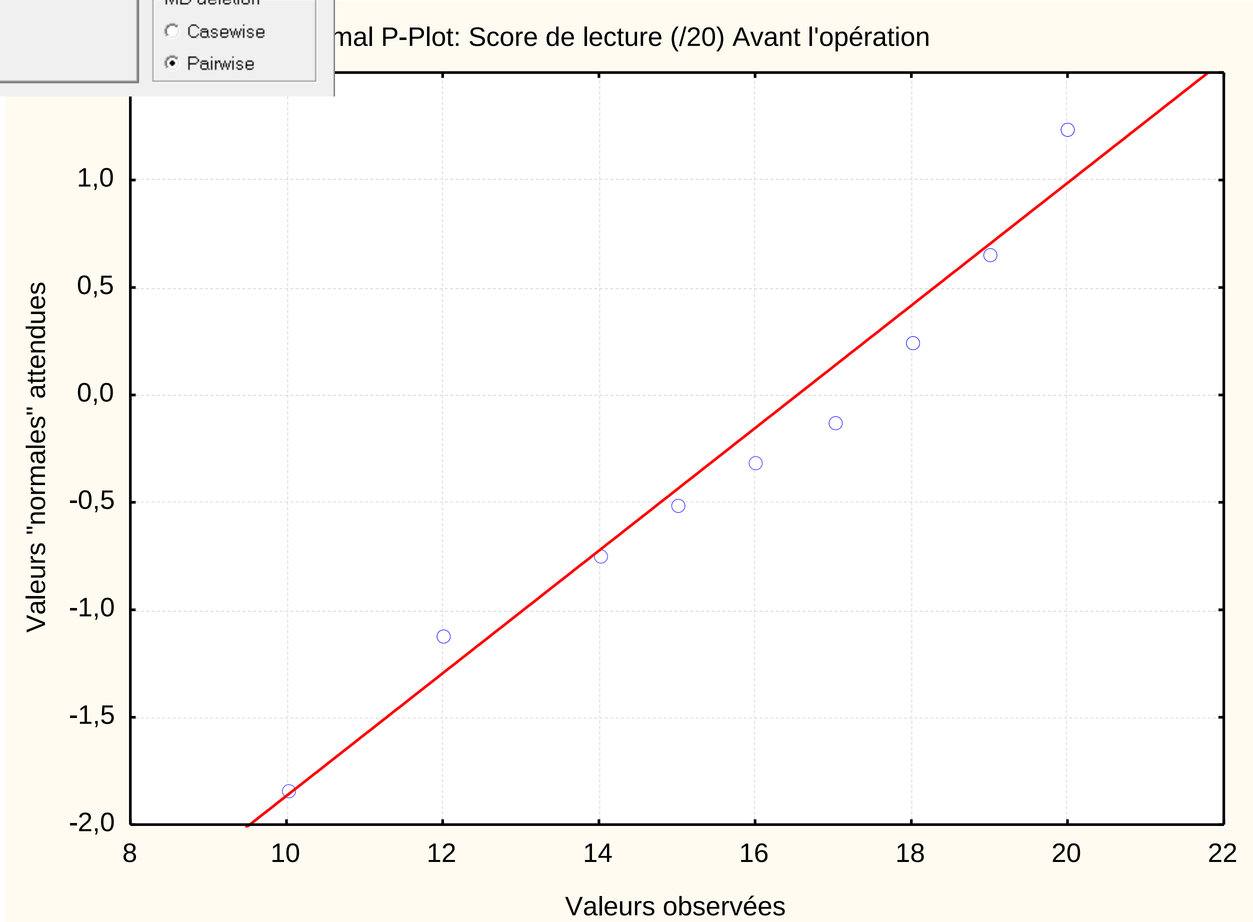
3- La distribution normale



Ce test de normalité suggère ici que la distribution des scores n’est pas normalement distribuée.



P-Plot



Si les valeurs observées s'éloignent trop de la droite, la distribution des données n'est probablement pas normale.

3- La distribution normale

❖ Formulation mathématique

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}(e)^{-(X-\mu)^2/2\sigma^2}$$

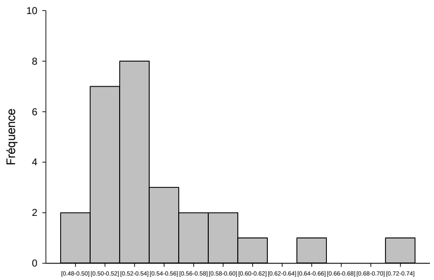
μ = moyenne
σ = Ecart-type
π = 3.1416
e = 2,7183

→ La forme de la distribution varie fortement avec l'écart type

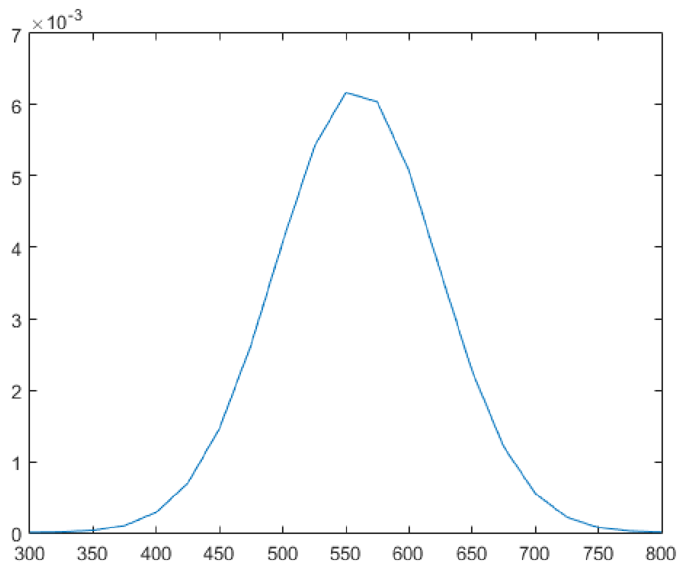
3- La distribution normale

❖ Formulation mathématique

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}(e)^{-(X-\mu)^2/2\sigma^2}$$



μ = 559 ms
σ = 64 ms

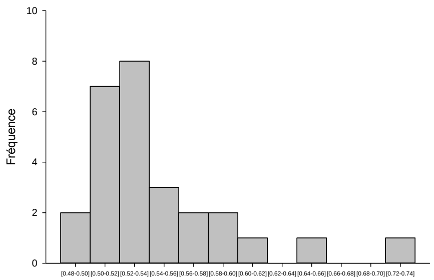


Courbe normale théorique en rentrant les paramètres donnés.

3-1 La distribution normale

3- La distribution normale

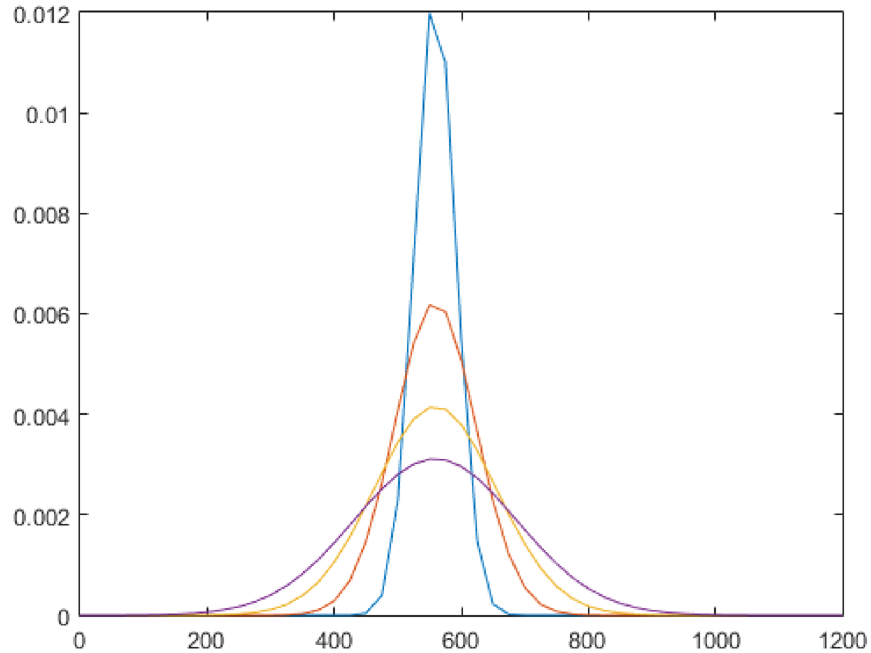
❖ Formulation mathématique



$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}(e)^{-(X-\mu)^2/2\sigma^2}$$

μ = 559 ms

$\sigma 1$ = 32 ms; $\sigma 2$ = 64 ms; $\sigma 3$ = 96 ms; $\sigma 4 = 128\ ms$



3-1 La distribution normale

3- La distribution normale

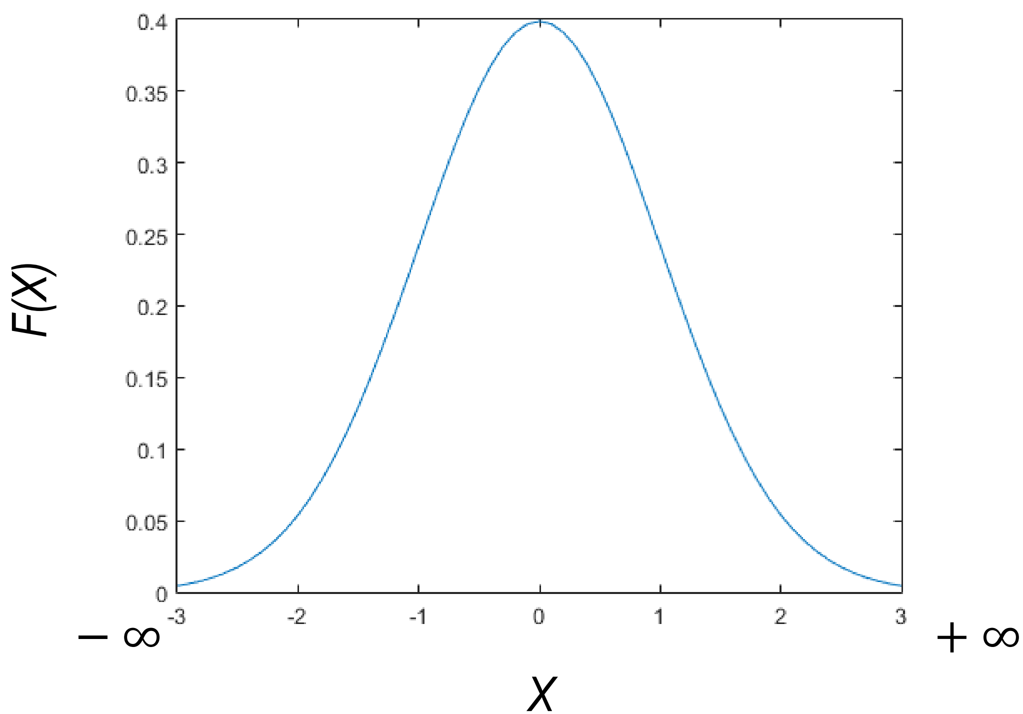
3-1 La distribution normale

3-2 Description normale centrée réduite

3- La distribution normale

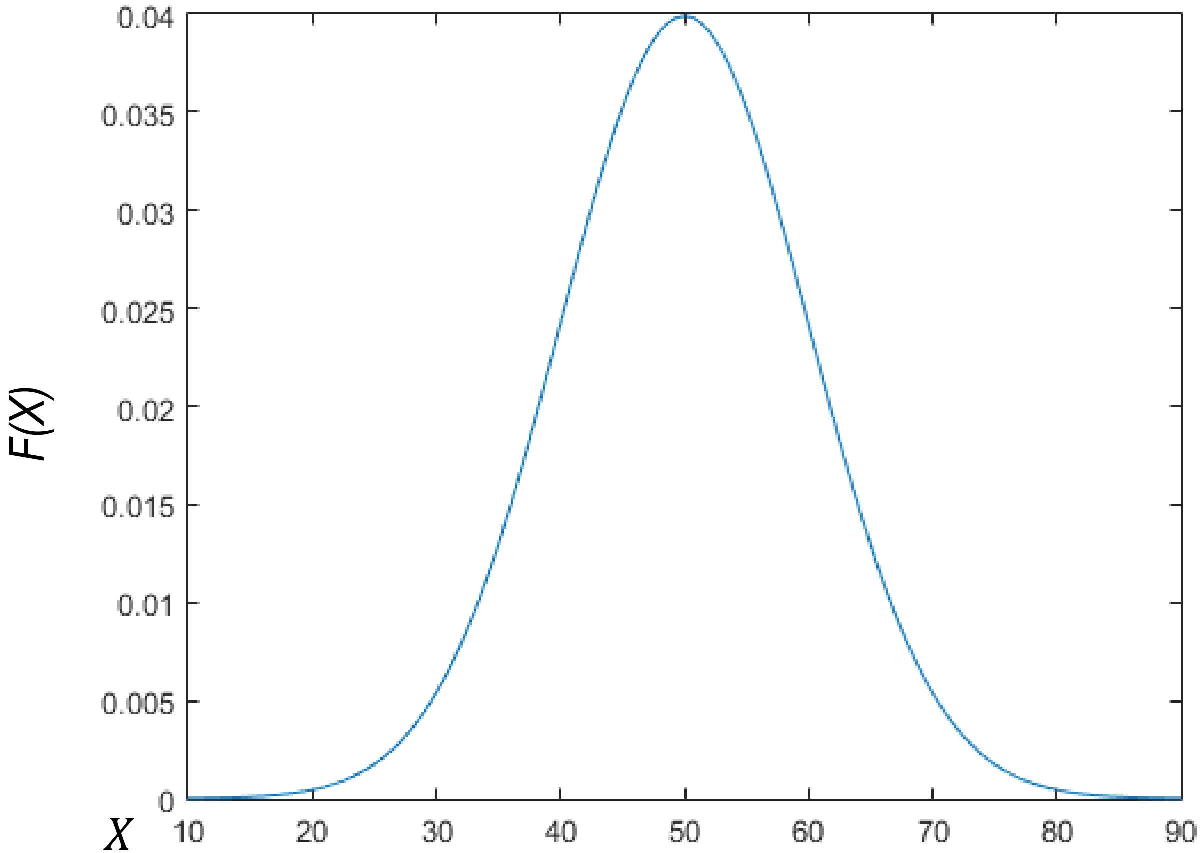
- La **distribution centrée réduite** est une distribution de moyenne 1 et d'écart-type 0. On la désigne souvent par *N*(0,1). On peut utiliser cette distribution pour faire n'importe quelle inférence puisque toute distribution normale peut-être transformée en distribution normale centrée réduite.

- La distribution N(0,1) est très utilise pour situer un individu (un score) par rapport à la distribution, et se prononcer sur la probabilité d'obtenir cette valeur dans la distribution.



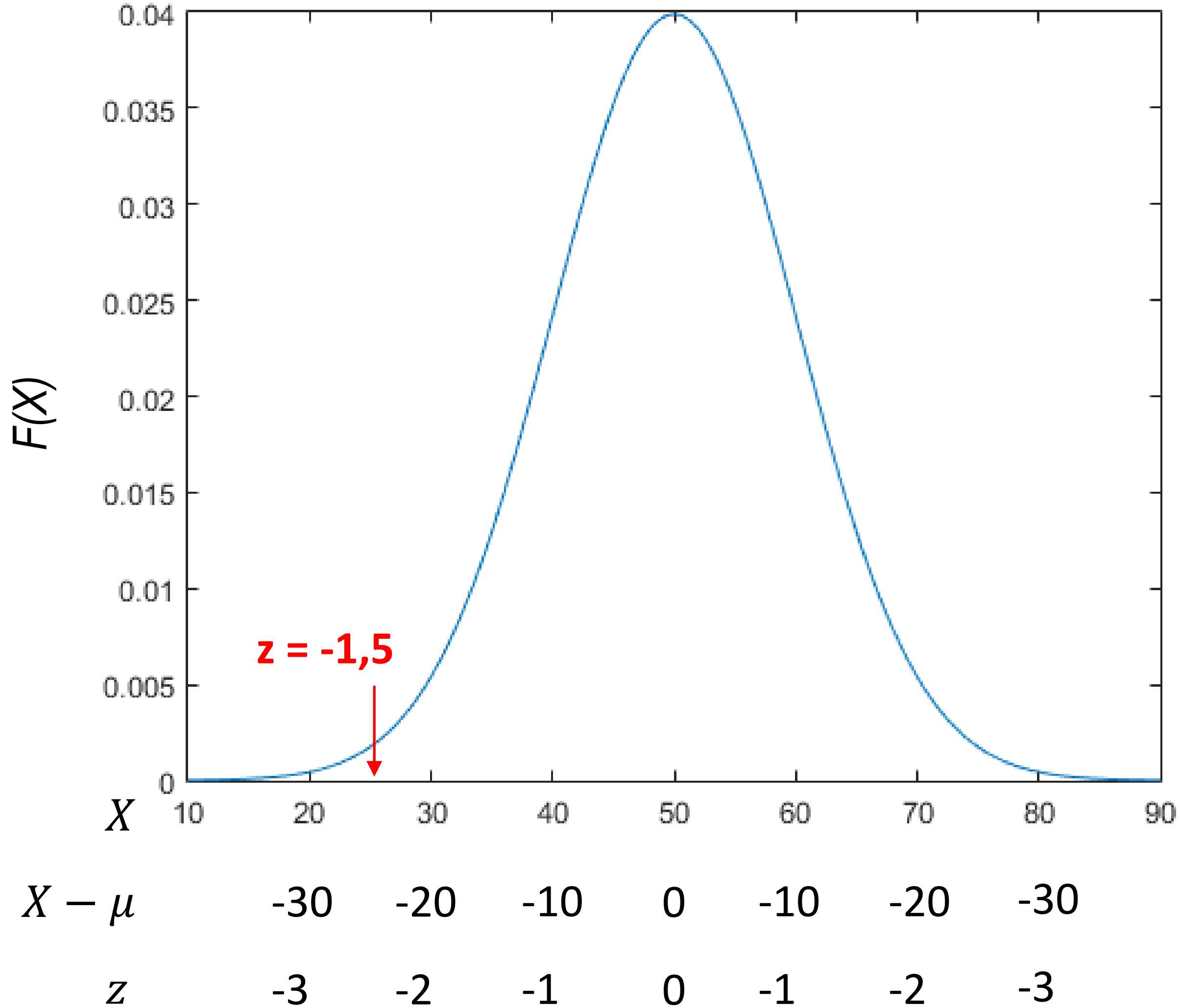
3- La distribution normale

μ = 50
σ = 10



X							
$X - \mu$	-30	-20	-10	0	-10	-20	-30
z	-3	-2	-1	0	-1	-2	-3

3- La distribution normale



Pour situer une valeur de distribution dans *N*(0,1), il suffit de procéder au calcul suivant :

$$z = \frac{X - \mu}{\sigma}$$

Par exemple, *X* = 35

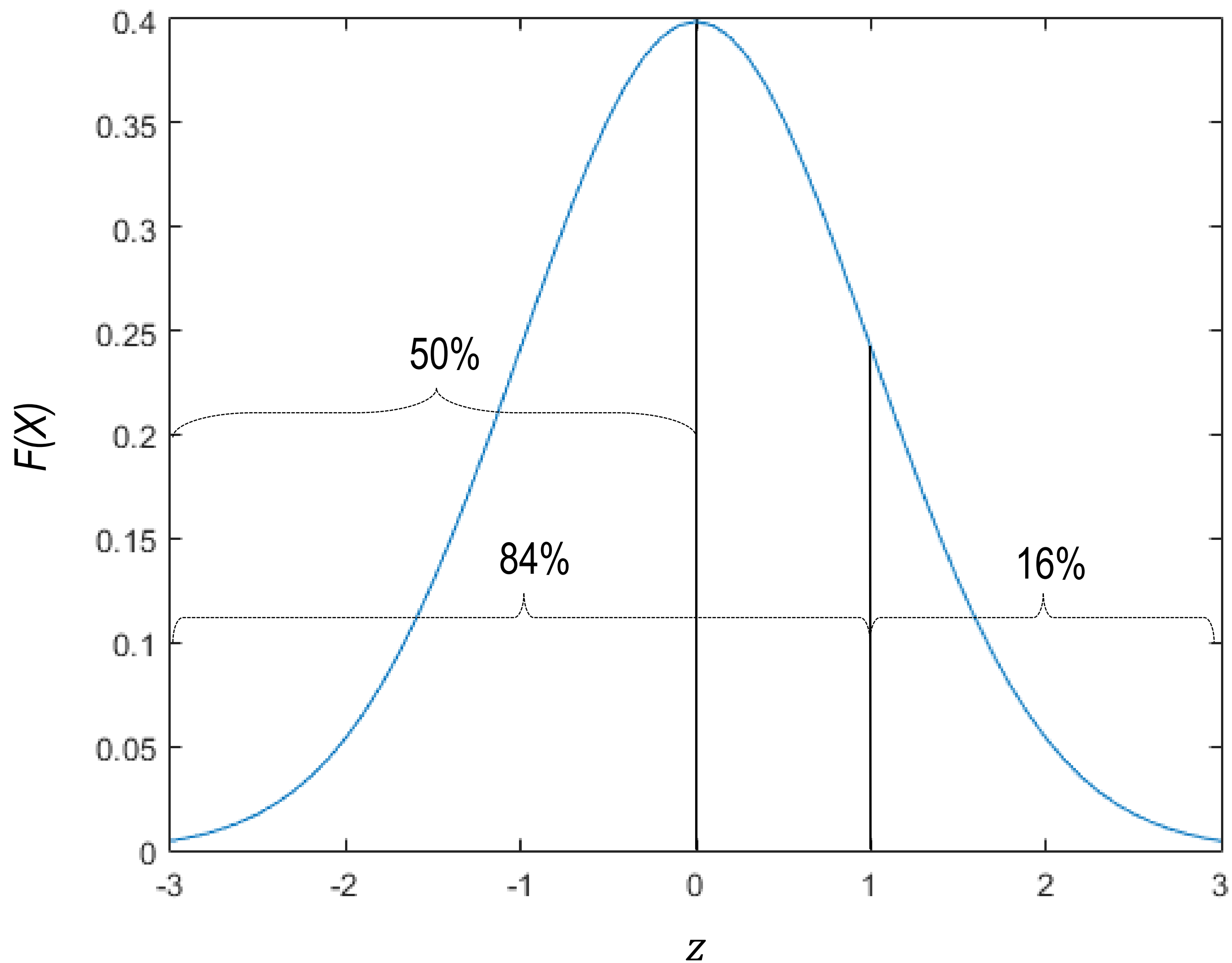
$$z = \frac{35 - 50}{10}$$

$$z = \frac{-15}{10}$$

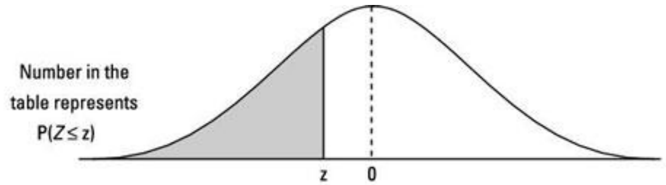
$$z = -1{,}5$$

3- La distribution normale

Pour déterminer la surface entre la moyenne et une valeur de z, il suffit de regarder dans les tables



3-2 La distribution centrée réduite



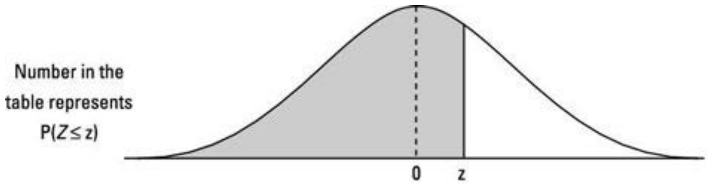
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Exemple:

z = - 0.37 signifie que l'aire grise sous la courbe contient 35.57% des individus/observations.

Z = - 1.96 signifie que l'aire grise sous la courbe contient seulement 2.5% des individus/observations

3-2 La distribution centrée réduite



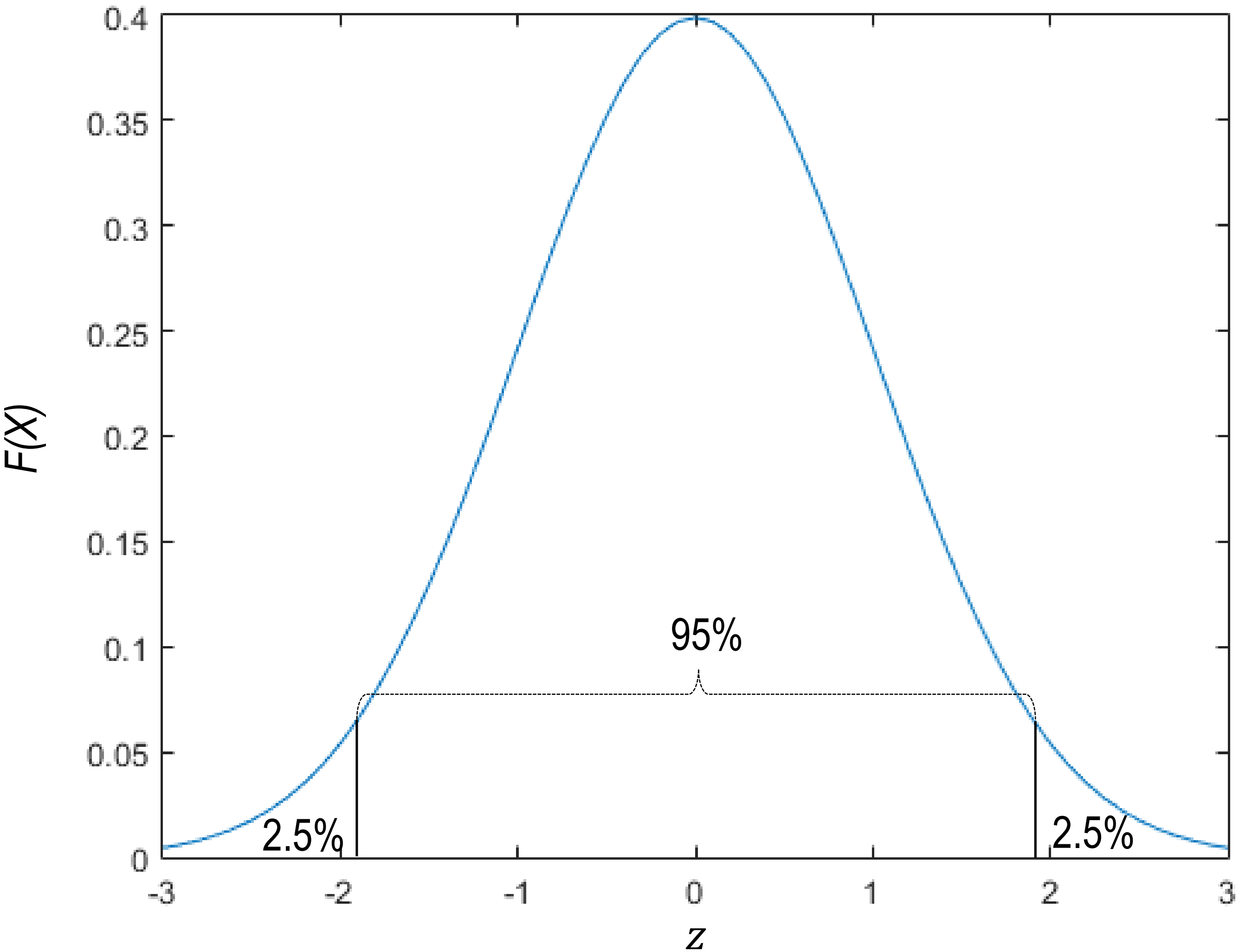
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Exemple:

z = 0.37 signifie que l’aire grise sous la courbe contient 64.53% des individus/observations.

Z = 1.96 signifie que l’aire grise sous la courbe contient seulement 97.50% des individus/observations

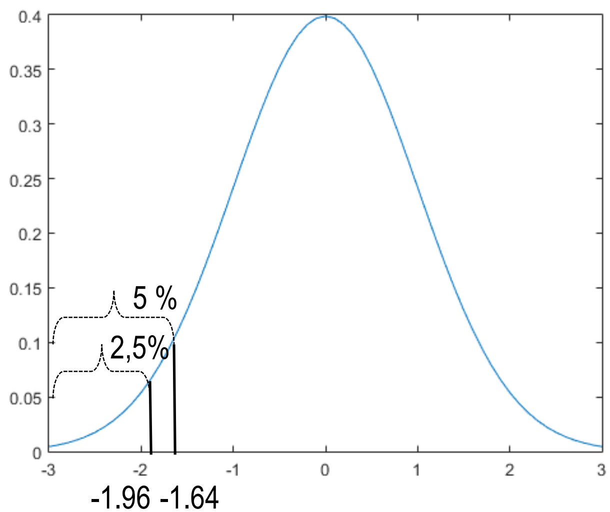
3- La distribution normale



Exemple : vous obtenez les résultats en dénomination de 200 sujets (moyenne = 60,4; ET = 10,3). Vous souhaitez savoir où se situe le score d'un sujet individuel qui a obtenu un score de 39. $z = \frac{X - \mu}{\sigma}$; $z = \frac{39 - 60,4}{10,3}$; $z =$ -2,07

3- La distribution normale

❖ Relation avec les probabilités



Supposez que vous avez les données normatives en lecture de 100 sujets normaux (moyenne 25, ET = 3,5).

Vous obtenez les scores de deux patients (19 et 15), et souhaitez savoir où ils se situent par rapport aux sujet normaux.

Vous allez calculer les scores.

$z = \frac{X - \mu}{\sigma}$; $z = \frac{19 - 25}{3,5}$; $z =$ -1,71 (Valeur du tableau P = 0,0436)

4,36 % de la population montre ce score

$z = \frac{X - \mu}{\sigma}$; $z = \frac{15 - 25}{3,5}$; $z =$ -2,85 (Valeur du tableau P = 0,0022)

0,2 % de la population montre ce score

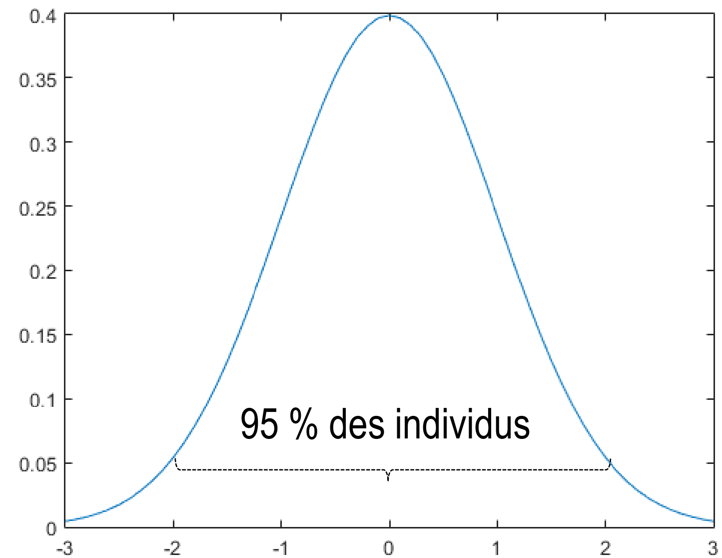
Ce qui équivaut à dire que vous avez seulement 4 chances sur 100, ou 0,2 chance sur 100 de vous tromper en disant que les deux patients montrent des performances anomales

3- La distribution normale

- Vous connaissez la distribution des scores à un test de dénomination d'objets, avec :

μ = 75
σ = 5

- Vous souhaitez calculer les bornes (les limites) entre lesquelles 95 % des scores sont distribués.



- Vous savez que $z = \frac{X - \mu}{\sigma}$

- Donc ± 1,96 = $= \frac{X - \mu}{\sigma} \rightarrow X - \mu$ = ± 1,96 σ $\rightarrow X = \mu$ ± 1,96 σ

- La limite inférieure est donc 75 – (1,96*5) = 65,2 et la limite supérieure 75 + (1,96*5) = 84,8

- <u>Interprétation</u> : « *la probabilité X d'observer un score entre 65,2 et 84,8 d'un sujet choisi aléatoirement est de 0.95 (95%)* »

- Il s'agit de l'intervalle de confiance ou limites de confiance

4- Hypothèses

4-1 Distribution d'échantillonnage

4-2 La théorie du test d'hypothèses

4-3 L'hypothèse nulle

4- Hypothèses

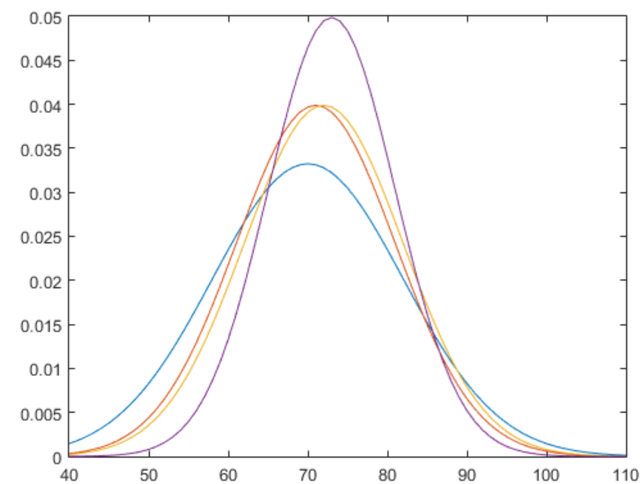
❖ <u>Variabilité due au hasard</u>

Supposez que vous voulez obtenir des normes pour un test évaluant la compréhension orale. Pour cela, vous tirez au sort des sujets de la population pour constituer votre échantillon.

Toutefois, il est très probable que, si vous sélectionnerez plusieurs échantillons, **la distribution des scores varient d'un échantillon à l'autre** – C'est-à-dire par exemple que la moyenne ne soit pas tout à fait la même dans les différents échantillons.

Cette variabilité attendue est appelée en statistique « ***variabilité due au hasard*** » ou parfois « ***erreur d'échantillonnage***. »

$Distri\ 1\!: \mu = 70, \sigma = 12$
$Distri\ 2\!: \mu = 61, \sigma = 10$
$Distri\ 3\!: \mu = 72, \sigma = 10$
$Distri\ 4\!: \mu = 73, \sigma = 8$

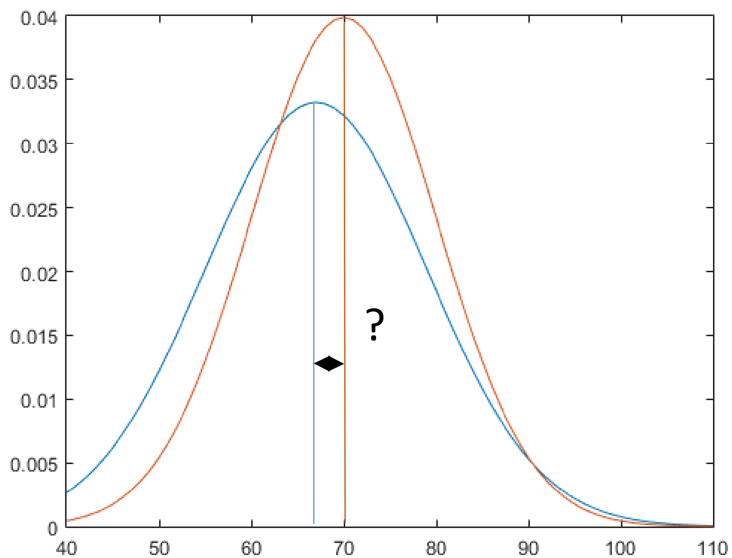


Variation de la distribution en fonction de l'échantillon

4- Hypothèses

❖ <u>Les variations sont-elles dues au hasard ?</u>

Supposez maintenant que vous avez les scores d'un échantillon de patients et les scores d'un échantillon de sujets normaux. La variation entre les deux distribution est-elle due au hasard ?



$Distri\ patients$: $\mu = 64, \sigma$ = 11
$Distri\ Sujets\ normaux$: $\mu = 70, \sigma$ = 10

Hypothèse 1 : la différence entre les deux moyennes est due au hasard ou à l'erreur d'échantillonnage.

Hypothèse 2: la différence entre les deux moyennes n'est pas due au hasard mais au fait que les patients ont une pathologie.

4- Hypothèses

❖ <u>Distribution d’échantillonnage</u>

Reprenons l’exemple précédent. La question posée est « *la différence entre les deux moyennes est-elle trop grande pour que l’on puisse l’attribuer à la chance ? »* Nous devons utiliser des **distributions d’échantillonnage**.

Le concept clé qui sous-tend les tests statistiques est la distribution d’échantillonnage d’une statistique. Sans celles-ci, les tests statistiques ne pourraient pas exister.

Ces distributions nous disent quelles valeurs nous pourrions (ou ne pourrions pas) nous attendre à obtenir pour une statistique prédéterminée et des conditions prédéfinies (à quelle différences de moyennes maximum faut-il atteindre ente les deux échantillons si les véritables moyennes des populations sont extraits les échantillons sont égales ?

La distribution d’échantillonnage d’une statistique **peut-être considérée comme la distribution des valeurs obtenues pour cette statistique sur un échantillonnage répété (un nombre illimité de fois)**.

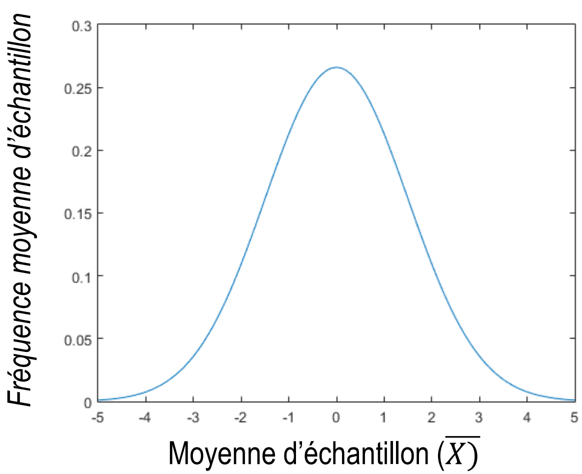
4- Hypothèses

4.1 Distribution d’échantillonnage

❖ <u>Distribution d’échantillonnage des différences entre moyennes</u>

- La distribution d’échantillonnage des différences entre les moyennes est la distribution des différences entre des moyennes d’un nombre infini d’échantillon aléatoire prélevé sous certaines conditions spécifiés (sous la condition par exemple que les vraies moyennes de la population sont égales).

- **Procédure théorique** : on prélève deux échantillons dans la population et ce de façon aléatoire, on calcule les différences de moyenne, puis on répète la procédure de façon infinie.



4- Hypothèses

4-1 Distribution d'échantillonnage
4-2 Test d'hypothèses
4-3 L'hypothèse nulle

4- Hypothèses

❖ Hypothèse de recherche

- Par exemple, nous souhaitons tester l'hypothèse que *les patients atteints d'une lésion dans l'aire de Broca se caractérisent par des difficultés au niveau des aspects expressifs du langage mises en évidence par le test de fluence phonologique* (dire tous les mots qui viennent à l'esprit et qui commencent par la lettre *P* en deux minutes).

- Pour ce faire, nous avons obtenu les données de deux échantillons ayant les mêmes caractéristiques sociodémographiques (âge, niveau d'éducation, sexe, *etc*.), se différenciant seulement par le fait que les sujets de l'un des échantillons ont une lésion dans l'aire de Broca.

4- Hypothèses

4-1 Distribution d’échantillonnage
4-2 Test d’hypothèses
4-3 L’hypothèse nulle
4-4 Statistiques de test

4- Hypothèses

❖ Hypothèse nulle

L'hypothèse nulle, **appelée H_0**, va se formuler de la façon suivante :

Il n'y a pas de différence entre les moyennes de deux échantillons

Ce qui revient à dire que :

$$\mu_{\text{Sujets normaux}} = \mu_{Patients}$$

Mais également que:

$$\mu_{\text{Sujets normau}x} + \mu_{Patients} = 0$$

4- Hypothèses

❖ Hypothèse nulle

Rationnelle d'utiliser le concept d'hypothèse nulle :

(1) Selon Fisher, nous ne pouvons jamais prouver le caractère purement exact d'une hypothèse, mais en démontrer son inexactitude.

Exemple : *Observer que 100 personnes ont deux mains ne suffit pas pour affirmer que toutes les personnes ont deux mains, mais trouver une personne qui a une main est suffisant pour mettre en branle l'hypothèse que toutes les personnes ont deux mains*.

(2) Ce type d'hypothèse est « pratique ».

Exemple : Vous souhaitez démontrer l'hypothèse que la moyenne des QI des orthophonistes est supérieure à 100. Quelle hypothèse allez-vous tester ? μ = 102 ; μ = 104; μ = 125 ?

Vous allez tester H_0 = 100 et éventuellement rejeter l'hypothèse.

4- Hypothèses

❖ <u>Hypothèse nulle a toujours sa contrepartie, appelée ***hypothèse alternative***</u>

Vous souhaitez démontrer l'hypothèse que la moyenne des QI des orthophonistes est supérieure à 100. Quelle hypothèse allez-vous tester ? μ = 102 ; μ = 104; μ = 125 ?

- Nous avons dit H_0 = 100

L'hypothèse alternative H_1 pourrait s'écrire H_1 : $\mu \neq 100$

Toutefois, celle-ci ne correspond pas à ce que vous vous voulez tester exactement. ***Il faut l'orienter***.

- L'hypothèse alternative H_1 doit s'écrire $\boldsymbol{H_1 : \mu > 100}$

4- Hypothèses

❖ <u>L'hypothèse alternative</u>

- Quand l'hypothèse alternative n'est pas orientée $\mu \neq 0$, on dit que l'hypothèse est **bilatérale** (*bicaudale* ou *bidirectionnelle*).

- Quand l'hypothèse est orientée $\mu < 0$ ou $\mu > 0$, on dit que l'hypothèse est unilatérale (*unicaudale* ou *unidirectionnelle*).

- Conséquence ; nous devons utiliser soit des tests unilatéraux, soit des tests bilatéraux. Dans le premier cas, il suffit de diviser la probabilité de signification par deux.

4- Hypothèses

4-1 Distribution d'échantillonnage
4-2 Test d'hypothèses
4-3 L'hypothèse nulle
4-4 Statistiques de test

4- Hypothèses

❖ Les statistiques de test

- Les statistiques que nous avons évoquées précédemment (ex: moyenne, médiane, écart-type, etc) sont des *statistiques d'échantillon*. Il existe d'autres catégories de statistiques, appelées **statistiques de test**, qui sont associées à des procédures statistiques bien précises et ont leurs propres distribution d'échantillonnage.

- Il s'agit par exemple du *t* de Student, du *F* de Fisher.

- Le t-test (ou t de Student) est souvent utiliser pour décider si la moyenne de deux échantillons a la même moyenne. $H_{0:}\mu_1 = \mu_2$

- La distribution d'échantillonnage théorique du *t*-test est simplement calculée en prévalant une quantité infinie de paires d'échantillon provenant de deux populations identiques, et en calculant le *t* pour chaque paire d'échantillons.

4- Hypothèses

4- Hypothèses

❖ Hypothèse sur une observation spécifique

- Supposons que vous vouliez comparer le score de lecture d'un patient atteint d'une lésion de la forme visuelle des mots (aire occipito-temporale basale) aux performances de sujets d'un échantillon provenant de la population normale.

- L'hypothèse nulle va se formuler de la façon suivante :

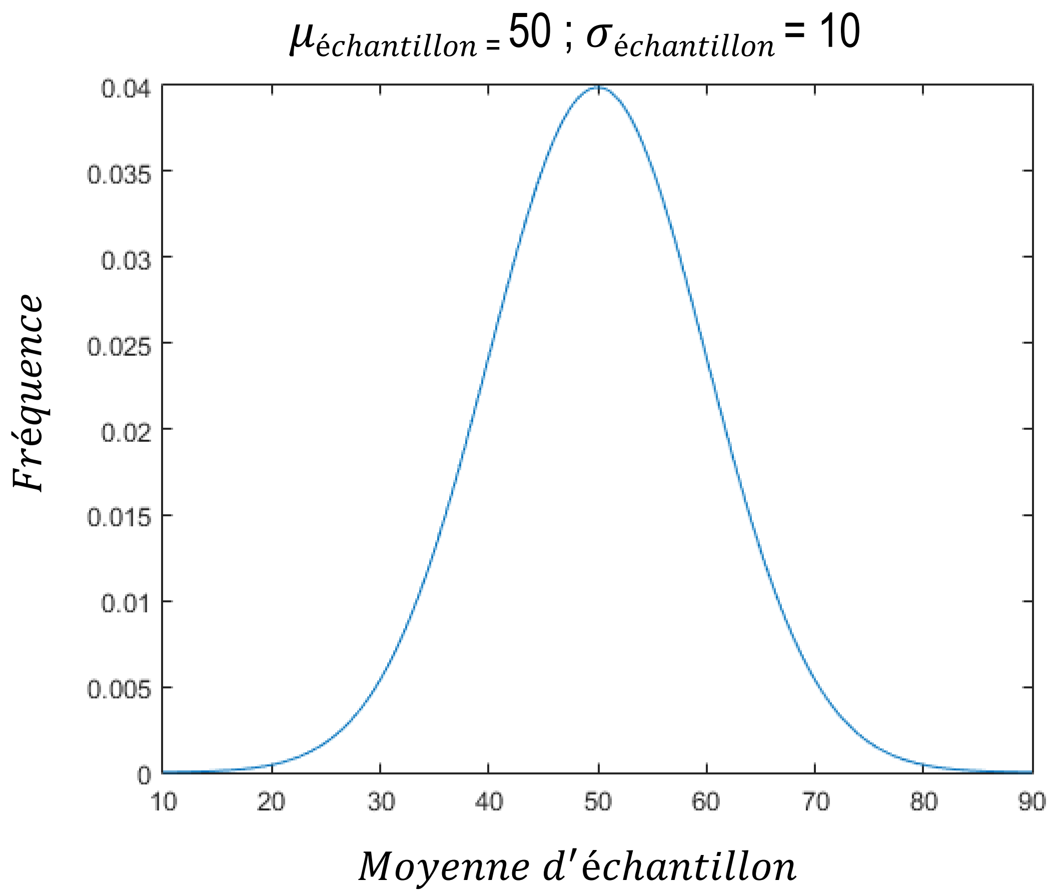
H_{0} : $Le\ score\ du\ patient\ provient\ de\ la\ population\ normale$ (le score du patient n'est pas différent de celui de la population normale)

- Grâce à la moyenne et à l'écart type de la distribution de l'échantillon, vous allez pouvoir calculer la probabilité d'obtenir un score aussi peu élevé.

=> Si la probabilité est très peu élevée, vous allez pouvoir rejeter H_{0} ;
=> Si la probabilité n'est pas faible, vous ne pouvez pas remettre en doute la validité de H_{0}.

4- Hypothèses

$$z = \frac{X - \mu}{\sigma}$$



$X_1 = 40;\quad z = \dfrac{40 - 50}{10} \quad z = -1$

$X_2 = 35;\quad z = \dfrac{35 - 50}{10} \quad z = -1.5$

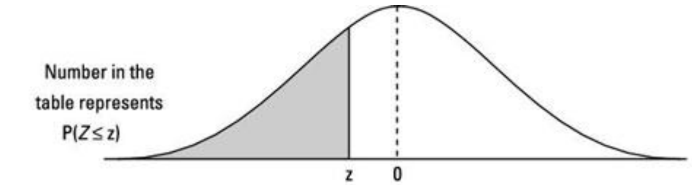
$X_3 = 20;\quad z = \dfrac{30 - 50}{10} \quad z = -2$

Vous regardez sur les tables

$z_1 = -1, P = 0{,}1587$

$z_2 = -1.5, P = 0{,}0668$

$z_3 = -2, P = 0{,}0228$



	z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
	−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
	−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
	−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
	−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
	−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
	−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
	−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
	−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
	−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
	−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
	−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
	−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
	−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
	−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
	−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
	−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
X_3	−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
	−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
	−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
	−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
	−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
X_2	−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
	−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
	−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
	−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
	−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
X_1	−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
	−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
	−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
	−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
	−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
	−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
	−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
	−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
	−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
	−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
	−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

4- Hypothèses

$$z_1 = -1, P = 0{,}1587$$

$$z_2 = -1.5, P = 0{,}0668$$

$$z_3 = -2, P = 0{,}0228$$

- En fonction de ces probabilités, vous allez devoir prendre une décision sur le rejet ou non de votre hypothèse nulle.

- Par convention, il y a rejet de H_0 lorsque la probabilité sous H_0 est inférieure ou égale à 0.5 ($P \leq .05$). Il s'agit du **seuil de rejet** ou encore appelé **seuil de signification**.

- Le seuil de rejet est défini de manière arbitraire. On juge d'une probabilité inférieure ou égale à .05 est suffisant pour rejeter l'hypothèse. Mais vous pourriez très bien utiliser un seuil plus conservateur de .01.

- Si l'on prend comme seuil de rejet .05, nous pouvons rejeter l'hypothèse nulle pour X_3 ; si .01, nous ne pouvons pas rejeter l'hypothèse $X_1 X_2 X_3$

5- Le test d’hypothèse appliqué aux moyennes

5-1 Distribution d’échantillonnage de la moyenne et théorème central limite

5-2 Le test t sur un échantillon (σ inconnu)

5-3 Test *t* pour échantillons appariés

5-4 Test *t* pour échantillons indépendants

5-5 Précautions d’application

5- Le test d'hypothèse appliqué aux moyennes

❖ Rappel

La **distribution d'échantillonnage d'une statistique** est la distribution des valeurs que nous nous attendrions à obtenir pour cette statistique si nous prélevions un nombre infini d'échantillons de la population en question et si nous calculions cette statistique sur chaque échantillon.

La distribution d'échantillon de la moyenne relève du même principe, et repose sur le **théorème central limite**.

« Etant donné une population ayant une moyenne μ et une variance σ², la distribution d'échantillonnage de la moyenne (la distribution des moyennes d'échantillon) aura une moyenne égale à μ (c'est-à-dire que $\mu_{\bar{X}} = \mu$), une variance de $\sigma^2_{\bar{X}}$ égale à $\frac{\sigma^2}{n}$ et un écart-type ($\sigma_{\bar{X}}$) égal à $\frac{\sigma}{\sqrt{n}}$, la distribution s'approchera de la distribution normale à mesure que n, l'effectif, augmente de 1. »

5- Le test d'hypothèse appliqué aux moyennes

❖ Nous souhaitons ici comparer les moyennes de deux échantillons dits **pairés** (ce qui correspond à des **mesures dîtes répétées**). Nous pouvons également dire:

- Echantillons liés;
- Echantillons corrélés;
- Echantillons appariés;
- Echantillons dépendants:

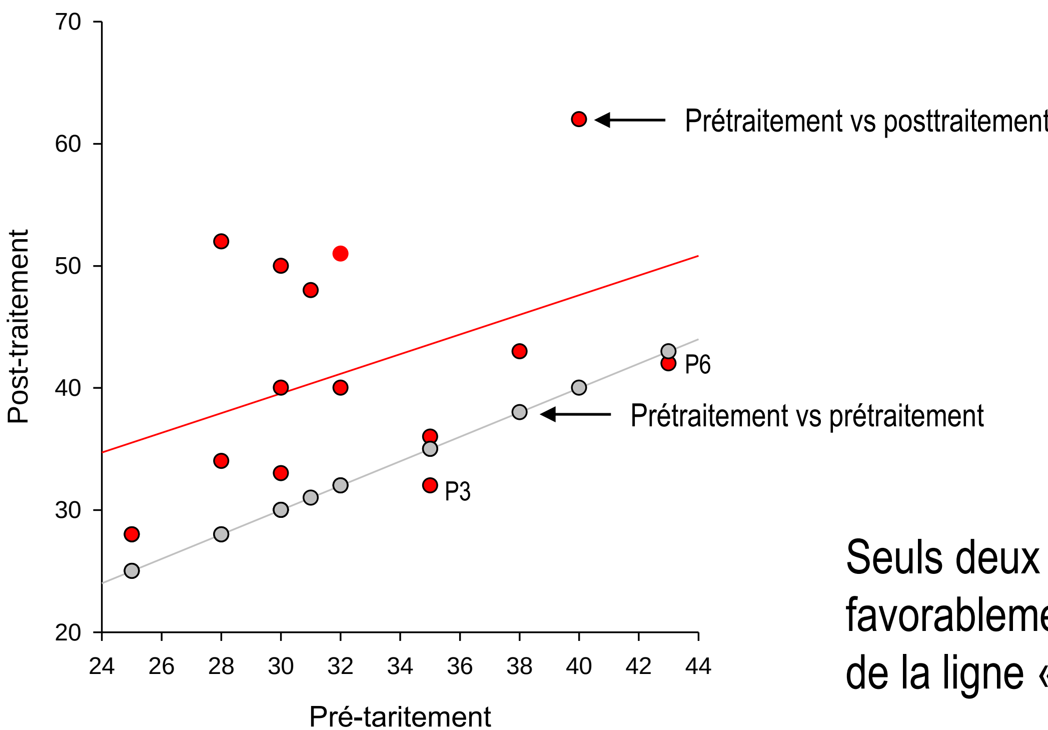
❖ Nous allons dans ce cas effectuer un *t*-test sur échantillons dépendants (ou mesures répétées.

❖ Prenons la situation expérimentale suivante : Vous souhaitez évaluer l'efficacité d'un programme de rééducation de l'accès au lexique. Pour mesurer l'accès au lexique, vous utilisez un test classique de dénomination que vous allez faire passer aux patients avant (prétraitement) et après la rééducation (*post-traitement*).

5- Le test d’hypothèse appliqué aux moyennes

Tableaux des données

Identification	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Moy.	ET.
Prétraitement	38	40	35	32	30	43	32	28	30	25	28	25	31	30	35	32,13	5,22
Posttraitement	43	62	32	40	40	42	51	52	33	28	34	28	48	50	36	41,27	9,84
Diff.	5	22	-3	8	10	-1	19	24	3	3	6	3	17	20	1	9,13	8,95



Seuls deux patients n’évaluent pas favorablement, ceux qui se situent en dessous de la ligne « grise »

5- Le test d'hypothèse appliqué aux moyennes

5.2 t-test pour échantillons pairés (ou appariés)

Tableaux des données

Identification	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Moy.	ET.
Prétraitement	38	40	35	32	30	43	32	28	30	25	28	25	31	30	35	32,13	5,22
Posttraitement	43	62	32	40	40	42	51	52	33	28	34	28	48	50	36	41,27	9,84
Diff.	5	22	-3	8	10	-1	19	24	3	3	6	3	17	20	1	9,13	8,95

- Cette différence de moyenne est-elle assez grande pour que l'on considère que la réhabilitation a eu un effet favorable sur la dénomination ?

$$H_0\colon\ \mu_{\text{Prétraitement}} = \mu_{Posttraitement}$$

- Toutefois, il est plus intéressant de s'intéresser directement aux scores de différences et ce pour chaque patient. Si la rééducation n'a aucun effet, la moyenne des scores des différences devrait être égale ou proche de 0. L'hypothèse nulle va donc s'écrire de la façon suivante:

$$H_0\colon\ \mu_{\text{D}} = \mu_{\text{Postraitement}} - \mu_{Prétraitement} = 0$$

5- Le test d'hypothèse appliqué aux moyennes

5.2 t-test pour échantillons pairés (ou appariés)

Tableaux des données

Identification	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**	**P15**	**Moy.**	**ET.**
Prétraitement	38	40	35	32	30	43	32	28	30	25	28	25	31	30	35	32,13	5,22
Posttraitement	43	62	32	40	40	42	51	52	33	28	34	28	48	50	36	41,27	9,84
Diff. (D)	5	22	-3	8	10	-1	19	24	3	3	6	3	17	20	1	9,13	8,95

❖ Calcul de *t*

$$t = \frac{\bar{D} - 0}{S_{\bar{D}}} = \frac{\bar{D} - 0}{\frac{S_D}{\sqrt{N}}}$$

Avec S_D: écart-type des différences

N : Nombre de différences

$\overline{D}$: Moyenne des différences

Dans notre exemple:

$$t = \frac{\bar{D} - 0}{S_{\bar{D}}} = \frac{\bar{D} - 0}{\frac{S_D}{\sqrt{N}}} = \frac{9{,}13 - 0}{\frac{8{,}95}{\sqrt{15}}} = \mathbf{3{,}95}$$

Le degré de liberté est toujours le nombre de paires moins 1
Donc 15-1 = 14

5- Le test d’hypothèse appliqué aux moyennes

5.2 t-test pour échantillons pairés (ou appariés)

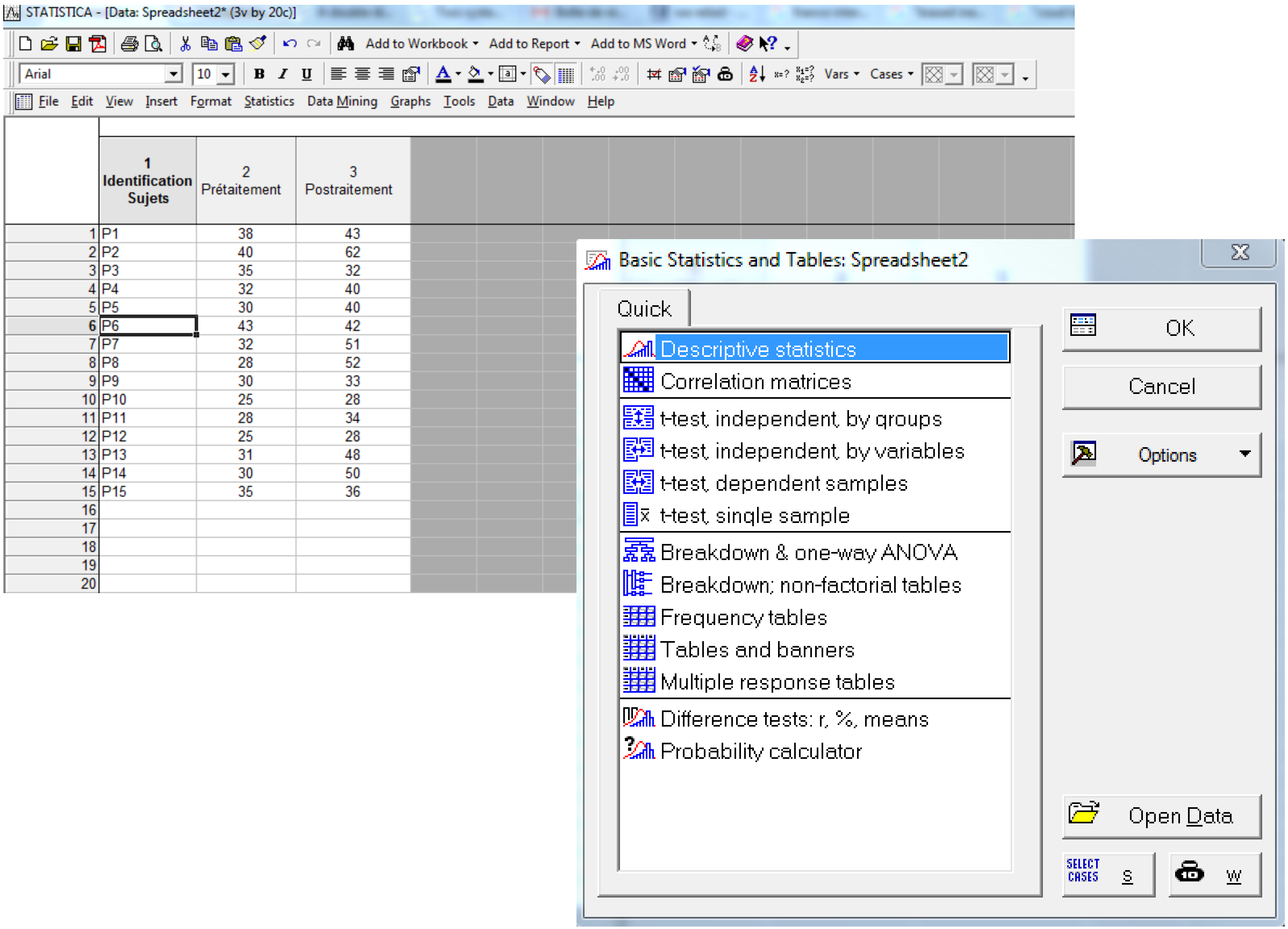
Degrees of freedom	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∝	1.282	1.645	1.960	2.326	2.576	3.291

$$t = \frac{\bar{D} - 0}{s_{\bar{D}}} = \frac{\bar{D} - 0}{\frac{s_D}{\sqrt{N}}} = \frac{9{,}13 - 0}{\frac{8{,}95}{\sqrt{15}}} = \mathbf{3{,}95}$$

- Si nous prenons un seuil de signification de α = .05, $t_{.05\,(16)} = \pm\, 2.145$.
- Etant donné que t = 3,95 dans notre exemple, nous pouvons rejeter l’hypothèse nulle et dire que les patients ont progressé de façon significative.
- Nous pouvons même rejeter l’hypothèse à un seuil inférieur à .01.
- Remarquez que si vous avions eu 4 sujets supplémentaires, nous aurions pu rejeter l’hypothèse à p < .001

5- Le test d’hypothèse appliqué aux moyennes

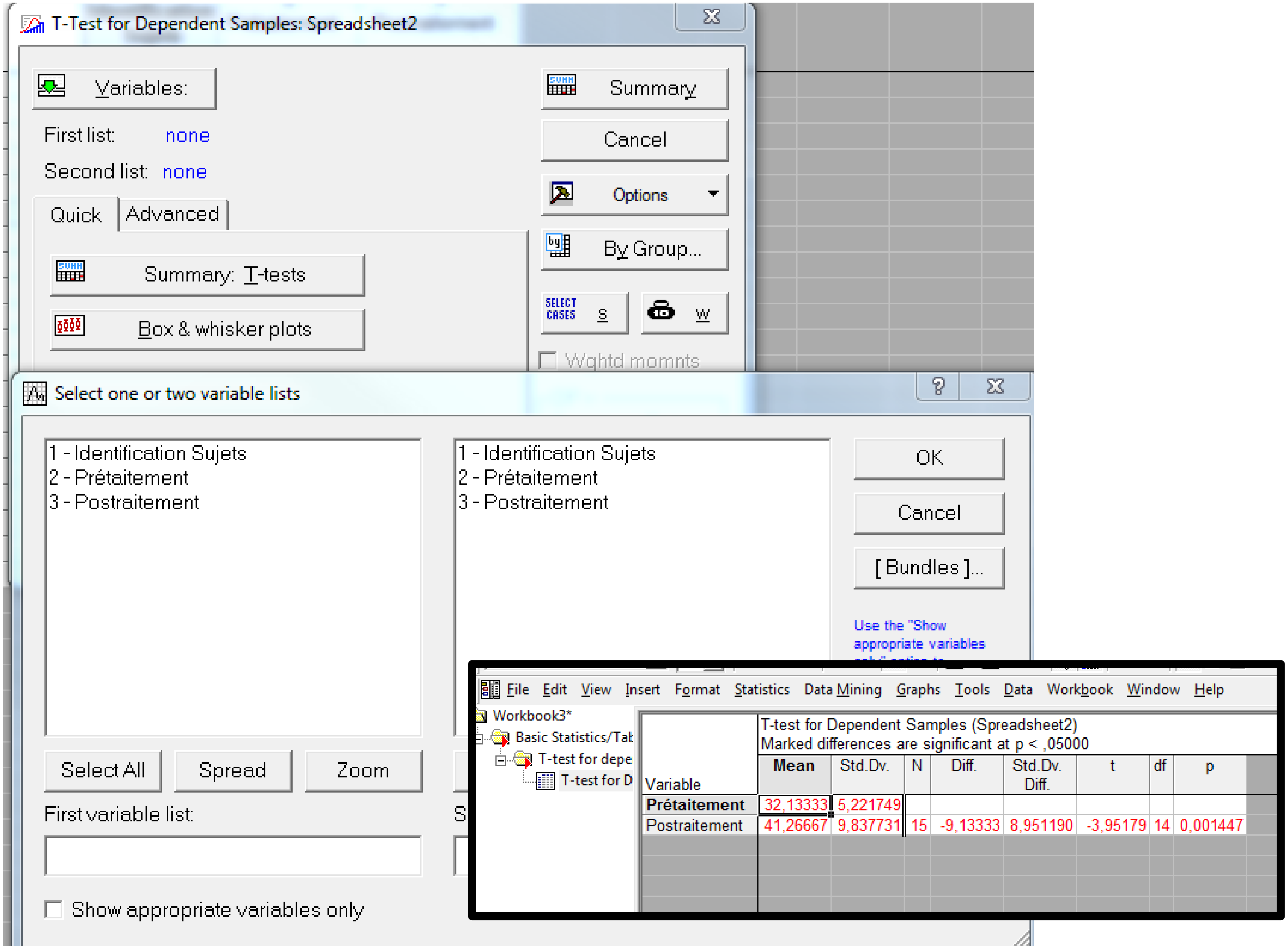
5.2 t-test pour échantillons pairés (ou appariés)



	1 Identification Sujets	2 Prétaitement	3 Postraitement
1	P1	38	43
2	P2	40	62
3	P3	35	32
4	P4	32	40
5	P5	30	40
6	P6	43	42
7	P7	32	51
8	P8	28	52
9	P9	30	33
10	P10	25	28
11	P11	28	34
12	P12	25	28
13	P13	31	48
14	P14	30	50
15	P15	35	36
16			
17			
18			
19			
20			

5- Le test d'hypothèse appliqué aux moyennes

5.2 t-test pour échantillons pairés (ou appariés)



5- Le test d'hypothèse appliqué aux moyennes

❖ Notation

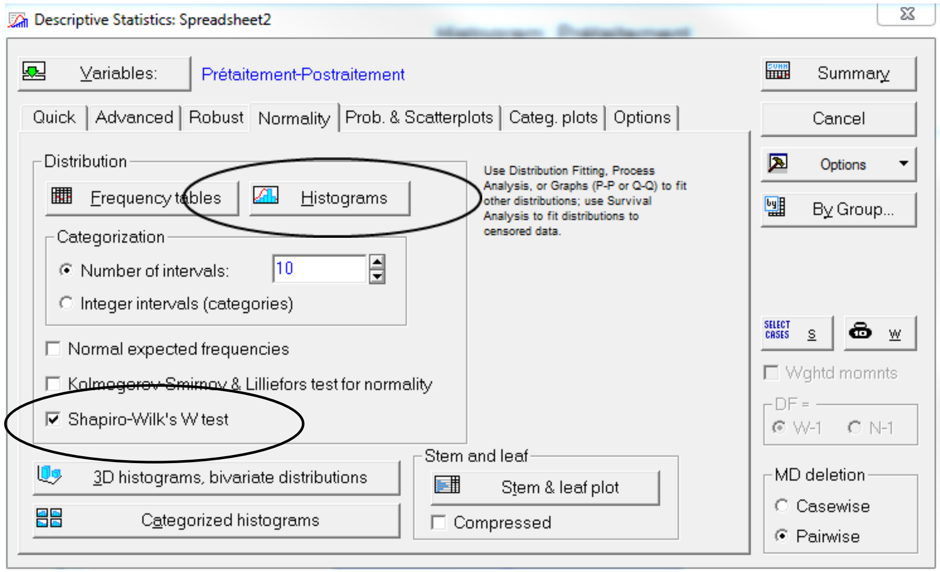
Dans les articles scientifiques (et dans votre mémoire), les résultats doivent être écrits de la manière suivante:

$t_{(14)} = 3{,}95$, $p < 0.01$ (ou < 0.005)

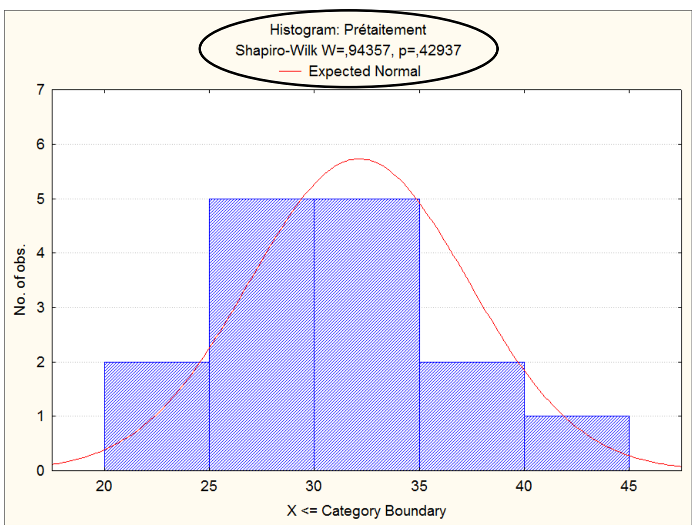
Remarque : étant donné que le t-test est basé sur les paramètres de la loi normale, nous aurions dû vérifier la normalité des données. Faisons maintenant, a *posteriori*.

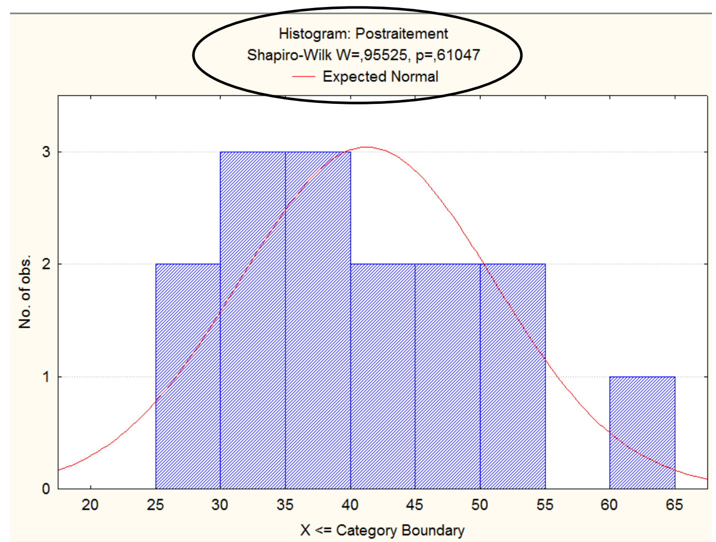
5- Le test d'hypothèse appliqué aux moyennes

5.2 t-test pour échantillons pairés (ou appariés)



Vous allez utiliser le test de Shapiro-Wilk's





5- Le test d'hypothèse appliqué aux moyennes

- Nous souhaitons ici comparer les moyennes de deux échantillons dits indépendants ce qui veut dire que les données proviennent de **deux groupes indépendants** et pas *des mêmes sujets*.

- Dans ce cas précis, nous allons utiliser un *t*-test pour échantillon indépendant dont la formulation mathématique est la suivante :

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

Les deux groupes ont le même nombre de sujets

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

Les deux groupes ont un nombre de sujets différents

Le *dl va s'obtenir en calculant :*

$n_1 + n_2 - 2$

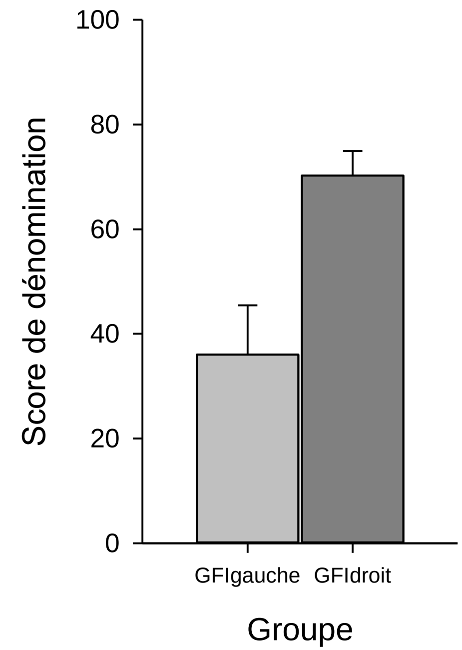
Remarque importante: la variance rentre en considération dans ces équations. Une des conditions d'applications de ce test est l'homogénéité des variances. Celle-ci doit donc être vérifiée.

5- Le test d’hypothèse appliqué aux moyennes

5.3 t-test pour échantillons indépendants

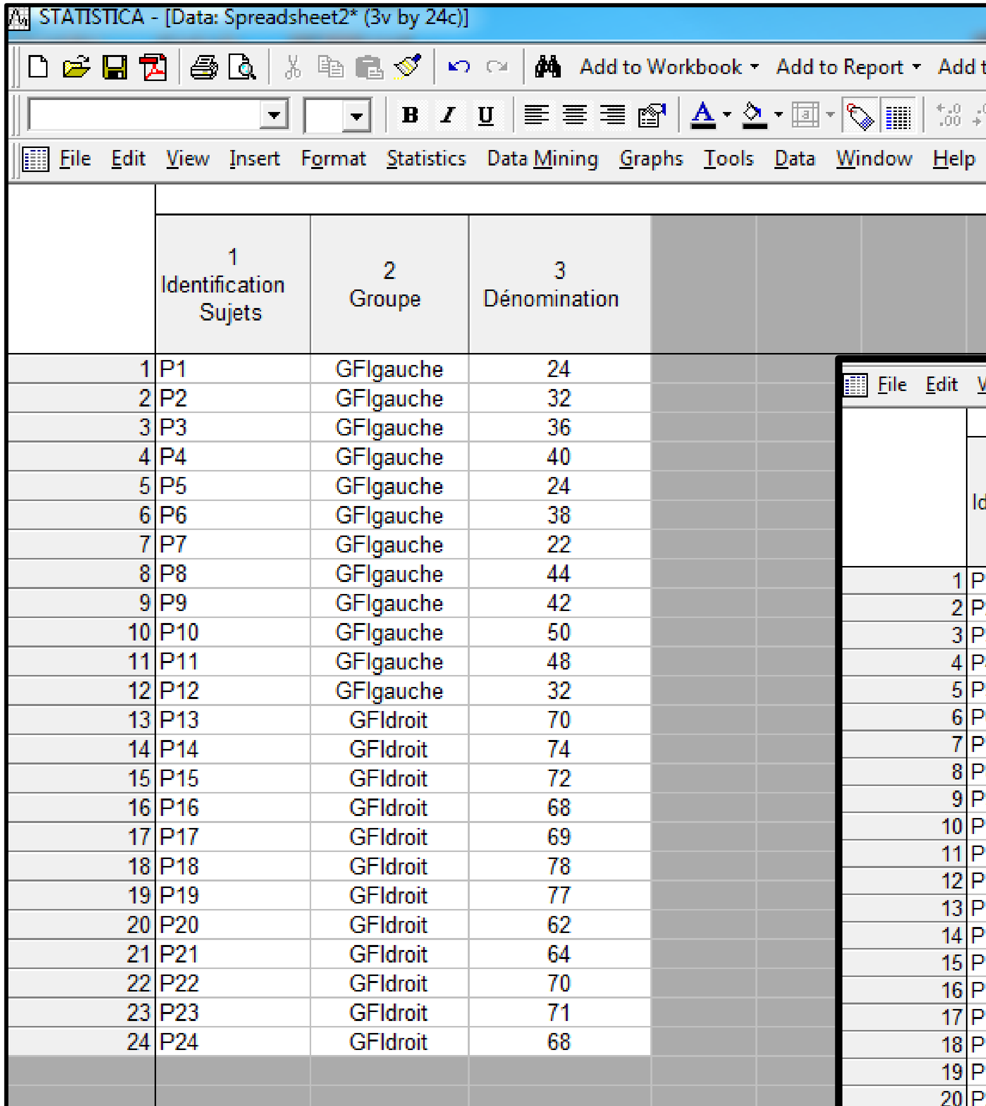
Identification	Groupe	Denomination
P1	GFIgauche	24
P2	GFIgauche	32
P3	GFIgauche	36
P4	GFIgauche	40
P5	GFIgauche	24
P6	GFIgauche	38
P7	GFIgauche	22
P8	GFIgauche	44
P9	GFIgauche	42
P10	GFIgauche	50
P11	GFIgauche	48
P12	GFIgauche	32
P13	GFIdroit	70
P14	GFIdroit	74
P15	GFIdroit	72
P16	GFIdroit	68
P17	GFIdroit	69
P18	GFIdroit	78
P19	GFIdroit	77
P20	GFIdroit	62
P21	GFIdroit	64
P22	GFIdroit	70
P23	GFIdroit	71
P24	GFIdroit	68

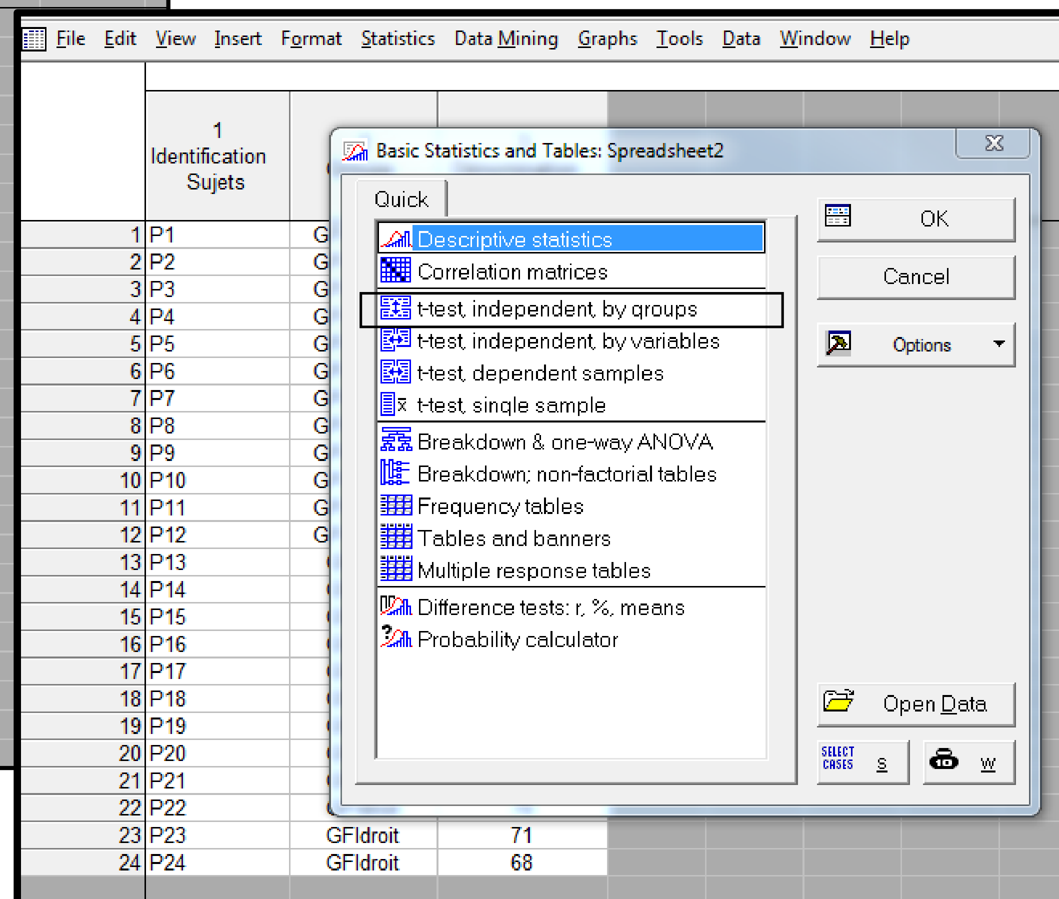
❖ Supposons que nous voulions comparer les performances en dénomination de patients avec une lésion du GFI gauche versus des patients atteints d’une lésion du GFI droit.

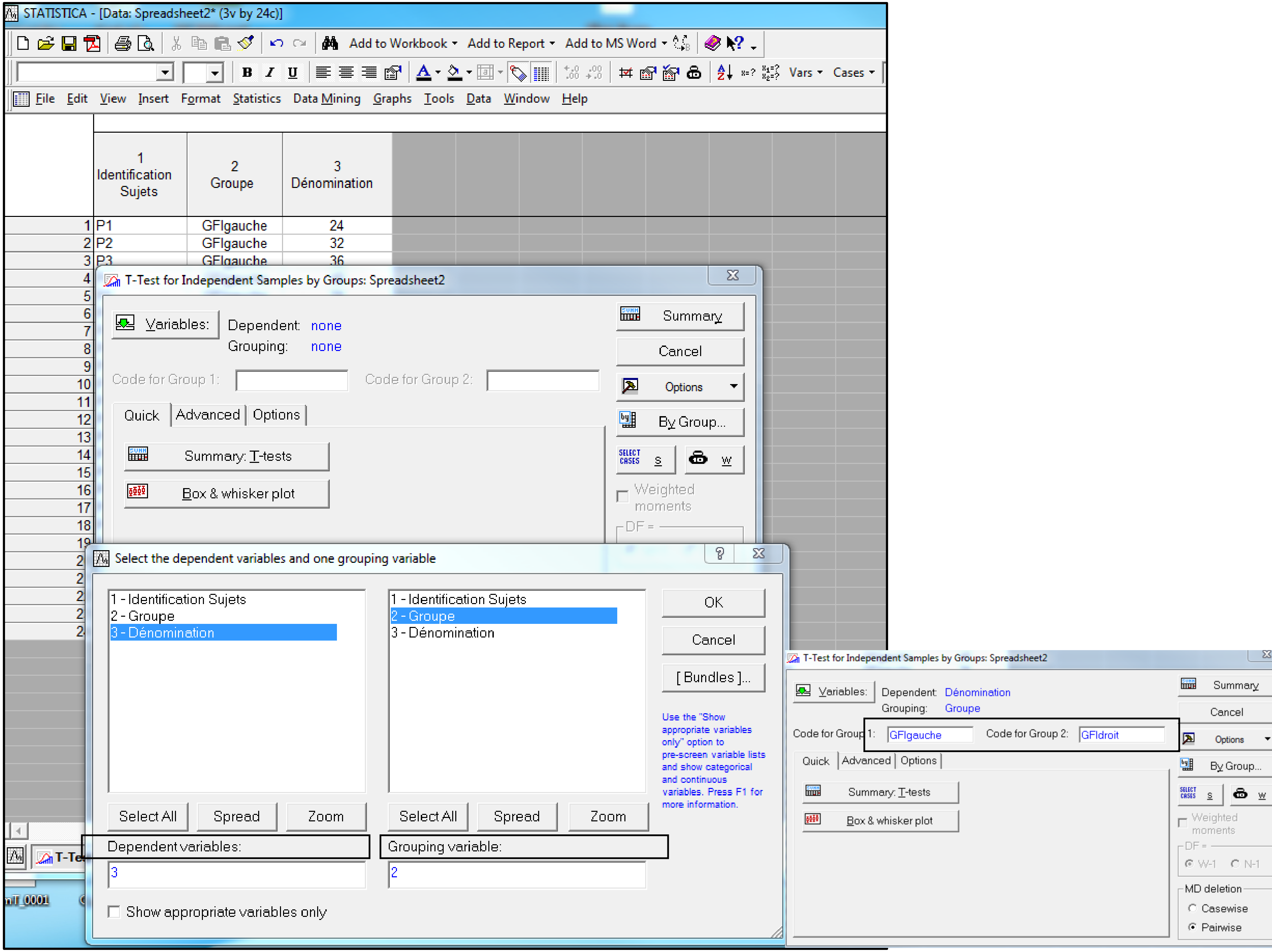


Que peut-on suspecter en regardant cet histogramme ?

5- Le test d'hypothèse appliqué aux moyennes






STATISTICA - [Data: Spreadsheet2* (3v by 24c)]
Add to Workbook
Add to Report
Add to MS Word
Vars
Cases
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
1 Identification Sujets
2 Groupe
3 Dénomination
1 P1 GFIgauche 24
2 P2 GFIgauche 32
3 P3 GFIgauche 36
T-Test for Independent Samples by Groups: Spreadsheet2
Variables:
Dependent: none
Grouping: none
Code for Group 1:
Code for Group 2:
Summary
Cancel
Options
By Group...
Quick
Advanced
Options
Summary: T-tests
Box & whisker plot
Weighted moments
DF =
Select the dependent variables and one grouping variable
1 - Identification Sujets
2 - Groupe
3 - Dénomination
1 - Identification Sujets
2 - Groupe
3 - Dénomination
OK
Cancel
[Bundles]...
Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.
Select All
Spread
Zoom
Select All
Spread
Zoom
Dependent variables:
3
Grouping variable:
2
Show appropriate variables only
T-Test for Independent Samples by Groups: Spreadsheet2
Variables:
Dependent: Dénomination
Grouping: Groupe
Code for Group 1: GFIgauche
Code for Group 2: GFIdroit
Summary
Cancel
Options
By Group...
Quick
Advanced
Options
Summary: T-tests
Box & whisker plot
Weighted moments
DF =
W-1
N-1
MD deletion
Casewise
Pairwise

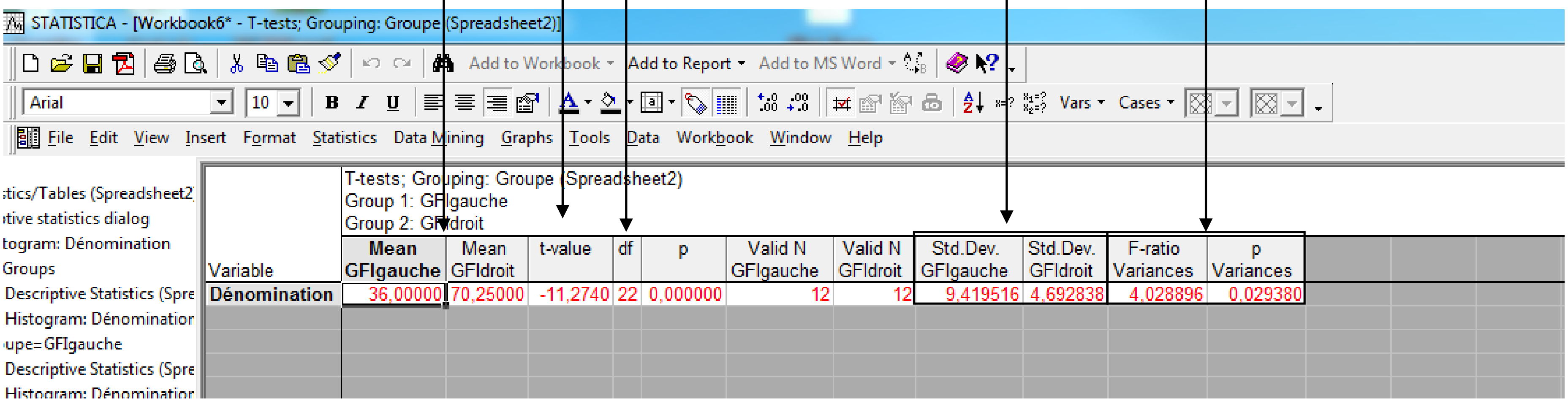
Moyenne des deux groupes

Valeur de *t*

Degré de liberté *(12+12-2)*

Les déviations standards sont éloignées

Le test du F-ratio démontre que les variances sont significativement différentes entre els deux groupes



Variable	Mean GFIgauche	Mean GFIdroit	t-value	df	p	Valid N GFIgauche	Valid N GFIdroit	Std.Dev. GFIgauche	Std.Dev. GFIdroit	F-ratio Variances	p Variances
Dénomination	36,00000	70,25000	-11,2740	22	0,000000	12	12	9,419516	4,692838	4,028896	0,029380

$t_{(22)} = -11,27,\ p < 0.0000001$

Remarque : dans ce cas précis, nous devrions plutôt utiliser des statistiques non-paramétriques du *fait de l'inhomogénéité des variances*.

6- Corrélation et régression linéaire

- 5-1 Le diagramme de dispersion
- 5-2 Le concept de covariance
- 6-3 Le coefficient de corrélation de Pearson
- 6-4 r ajusté
- 6-5 Droite de régression
- 6-6 la signification de r²
- 6-7 Test de signification de r et b
- 6-8 Les facteurs influençant la corrélation

6- Corrélation et régression

6.1 Diagramme de dispersion

❖ Ici, nous souhaitons évaluer la relation qu'il existe entre deux variables X et Y.

❖ Qu'est ce qui différencie une corrélation d'une régression simple :

- Dans la corrélation, X et Y sont des ***variables aléatoires*** (une variable qui échappe au contrôle de l'expérimentateur)

- Dans la régression, les modalités de X sont fixés par l'expérimentateur (c'est-à-dire manipulé par l'expérimentateur)

❖ Une façon plus simple de distinguer ces deux types d'analyse :

- On utilisera la corrélation si nous souhaitons simplement étudier la force du lien entre les deux variables X et Y;

- On utilisera la régression simple si nous souhaitons prédire les valeurs de Y sur la base de X;

6- Corrélation et régression

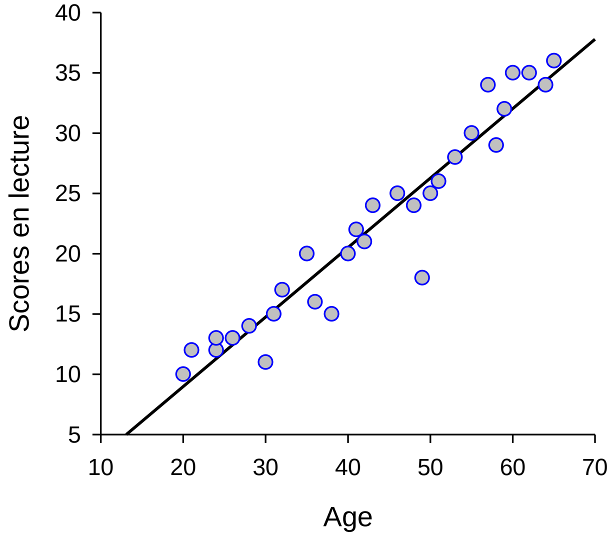
6.1 Diagramme de dispersion

- Le diagramme de dispersion représente chaque sujet expérimental inclus dans l'étude par un point dans un espace bidimensionnel. Les cordonnées de ces points (*Xi*, *Y*i) sont les scores obtenus par l'individu pour les variables X et Y, respectivement.

- *Important* : on suppose qu'il **existe une relation linéaire**.

- La variable qui sert de prédicteur sera toujours mis en abscisse alors que la variable qui sert de critère (c'est-à-dire c'est qui doit être prédite, en ordonnée).

- Dans le contexte d'une corrélation, l'emplacement de X ou Y importe peu.



6- Corrélation et régression

6.1 Diagramme de dispersion

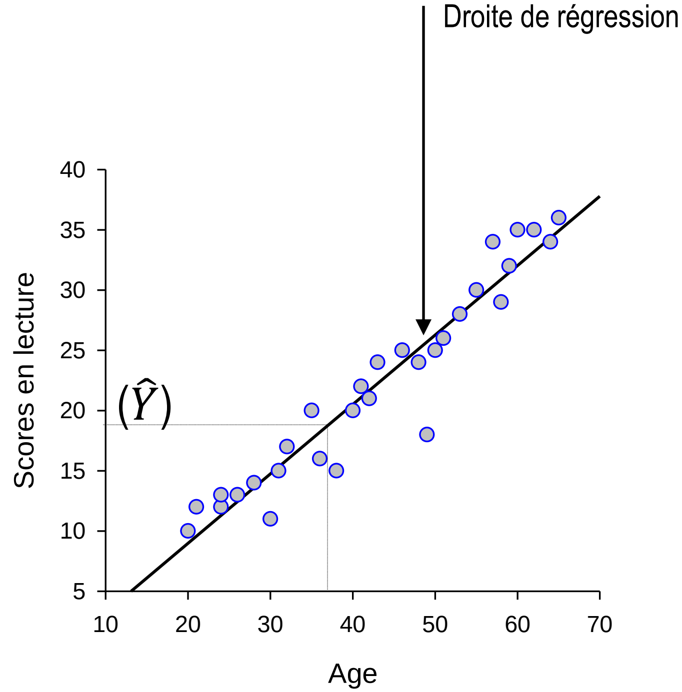
- La **droite régression** est la droite qui s'ajuste le mieux aux données, elle représente la meilleure prédiction de Y_i en fonction de X_i

- « Y chapeau » ($\hat{Y}$) représente la meilleure prédiction de Y pour un X donné.

- Le *degré auquel les points se regroupent autour de la droite de régression* est lié à la corrélation ***r*** entre X et Y.

- Si les points étaient tous alignés sur la droite la corrélation serait parfaite et équivaudrait à + 1

- Le coefficient de corrélation est toujours situé en -1 et +1



6- Corrélation et régression linéaire

5-1 Le diagramme de dispersion

5-2 Le concept de covariance

6-3 Le coefficient de corrélation de Pearson

6-4 r ajusté

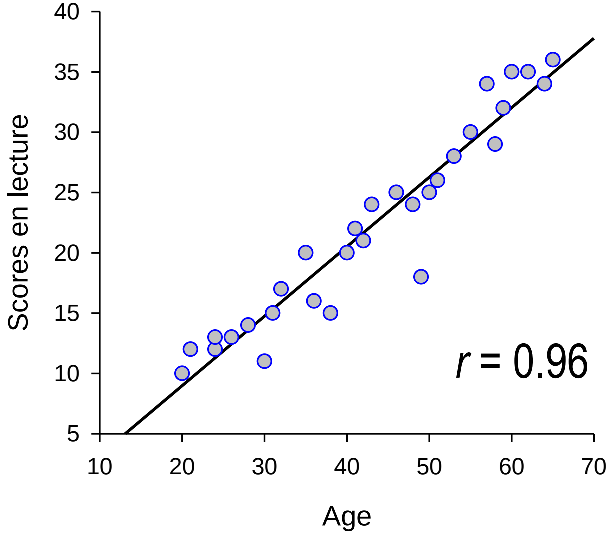
6-5 Droite de régression

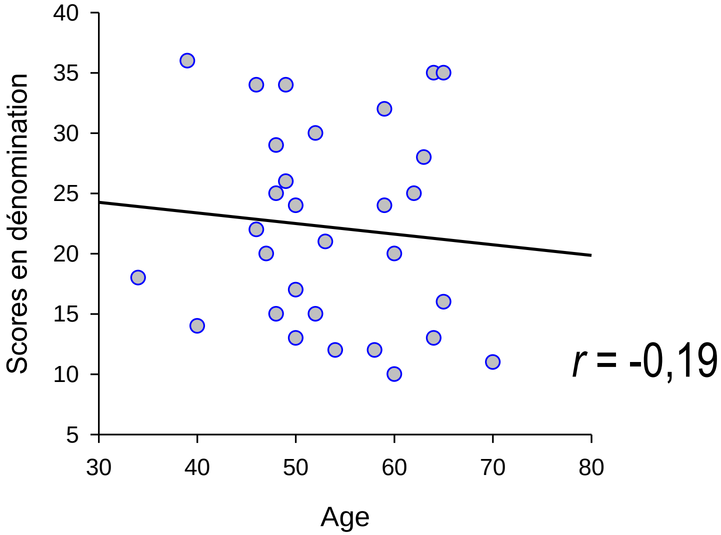
6-6 la signification de r^2

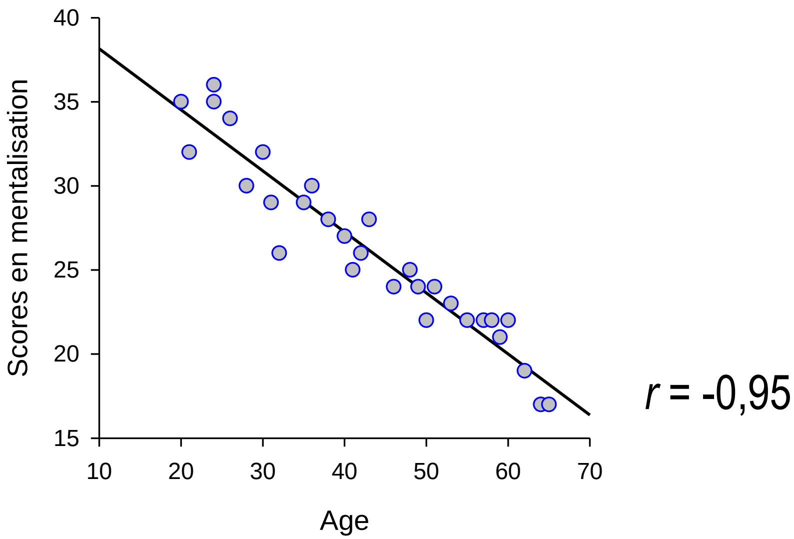
6-7 Test de signification de r et b

6-8 Les facteurs influençant la corrélation

6- Corrélation et régression







6.2 Covariance

6- Corrélation et régression

- Pour calculer une corrélation, nous avons besoin d'une statistique appelée la **covariance** (cov_{XY}). La covariance est un nombre qui reflète le degré auquel deux variables varient ensemble.

- D'un point de vue formel :

$$COV_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N-1} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{N-1}$$

6- Corrélation et régression

6.2 Covraiance

n sujets	Age (X)	Denomination (Y)	$(X-\overline{X})$	$(Y-\overline{Y})$	$(X-\overline{X})(Y-\overline{Y})$
1	20	10	-22,93	-12,20	279,79
2	21	12	-21,93	-10,20	223,72
3	24	12	-18,93	-10,20	193,12
4	24	13	-18,93	-9,20	174,19
5	26	13	-16,93	-9,20	155,79
6	28	14	-14,93	-8,20	122,45
7	30	11	-12,93	-11,20	144,85
8	31	15	-11,93	-7,20	85,92
9	32	17	-10,93	-5,20	56,85
10	35	20	-7,93	-2,20	17,45
11	36	16	-6,93	-6,20	42,99
12	38	15	-4,93	-7,20	35,52
13	40	20	-2,93	-2,20	6,45
14	41	22	-1,93	-0,20	0,39
15	42	21	-0,93	-1,20	1,12
16	43	24	0,07	1,80	0,12
17	46	25	3,07	2,80	8,59
18	48	24	5,07	1,80	9,12
19	49	18	6,07	-4,20	-25,48
20	50	25	7,07	2,80	19,79
21	51	26	8,07	3,80	30,65
22	53	28	10,07	5,80	58,39
23	55	30	12,07	7,80	94,12
24	57	34	14,07	11,80	165,99
25	58	29	15,07	6,80	102,45
26	59	32	16,07	9,80	157,45
27	60	35	17,07	12,80	218,45
28	62	35	19,07	12,80	244,05
29	64	34	21,07	11,80	248,59
30	65	36	22,07	13,80	304,52
Moyenne	42,93	22,20			3177,40
ET	13,79	8,28			
					Somme

$$COV_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N-1}$$

$$COV_{XY} = \frac{3177,40}{30-1} = 109,56$$

6- Corrélation et régression linéaire

5-1 Le diagramme de dispersion

5-2 Le concept de covariance

6-3 Le coefficient de corrélation de Pearson

6-4 r ajusté

6-5 Droite de régression

6-6 la signification de r²

6-7 Test de signification de r et b

6-8 Les facteurs influençant la corrélation

6- Corrélation et régression

6.3Calcul du coefficient de corrélation

- Le calcul du <u>coefficient de Pearson </u>*r* s'effectue de la façon suivante:

$$r = \frac{COV_{YX}}{S_X S_Y}$$

Ce coefficient est toujours situé entre 1 et -1:

- Plus on tend vers 1 ou -1, plus la relation est forte ;
- Plus on s'approche de 0, plus la relation est faible ;
- le signe + ou le signe – indique le sens de la relation ;

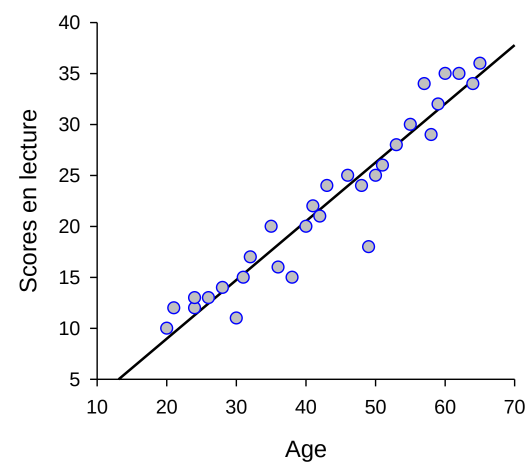
6- Corrélation et régression

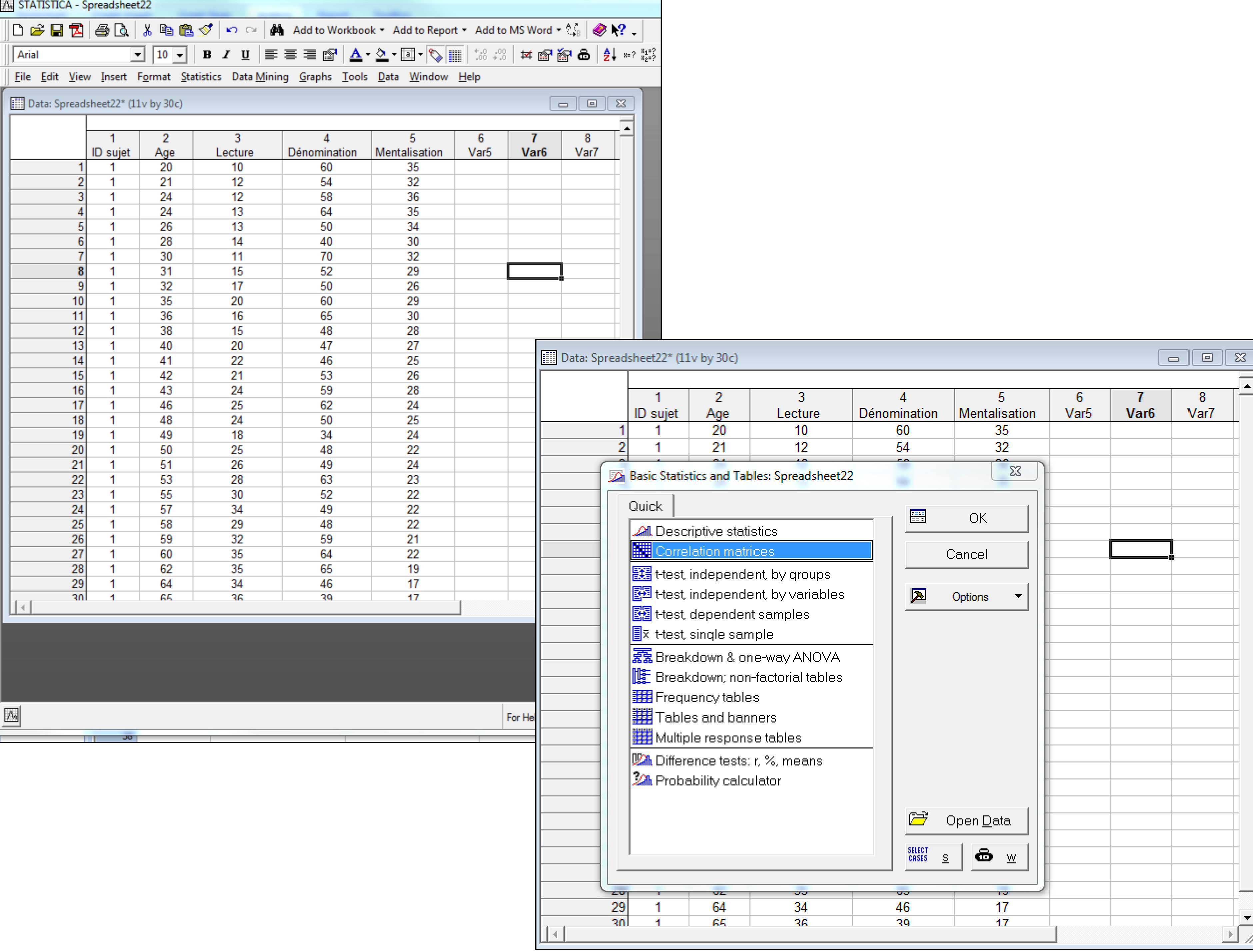
6.3Calcul du coefficient de corrélation

n sujets	Age (X)	Denomination (Y)	$(X-\overline{X})$	$(Y-\overline{Y})$	$(X-\overline{X})(Y-\overline{Y})$
1	20	10	-22,93	-12,20	279,79
2	21	12	-21,93	-10,20	223,72
3	24	12	-18,93	-10,20	193,12
4	24	13	-18,93	-9,20	174,19
5	26	13	-16,93	-9,20	155,79
6	28	14	-14,93	-8,20	122,45
7	30	11	-12,93	-11,20	144,85
8	31	15	-11,93	-7,20	85,92
9	32	17	-10,93	-5,20	56,85
10	35	20	-7,93	-2,20	17,45
11	36	16	-6,93	-6,20	42,99
12	38	15	-4,93	-7,20	35,52
13	40	20	-2,93	-2,20	6,45
14	41	22	-1,93	-0,20	0,39
15	42	21	-0,93	-1,20	1,12
16	43	24	0,07	1,80	0,12
17	46	25	3,07	2,80	8,59
18	48	24	5,07	1,80	9,12
19	49	18	6,07	-4,20	-25,48
20	50	25	7,07	2,80	19,79
21	51	26	8,07	3,80	30,65
22	53	28	10,07	5,80	58,39
23	55	30	12,07	7,80	94,12
24	57	34	14,07	11,80	165,99
25	58	29	15,07	6,80	102,45
26	59	32	16,07	9,80	157,45
27	60	35	17,07	12,80	218,45
28	62	35	19,07	12,80	244,05
29	64	34	21,07	11,80	248,59
30	65	36	22,07	13,80	304,52
Moyenne	42,93	22,20			3177,40
ET	13,79	8,28			Somme

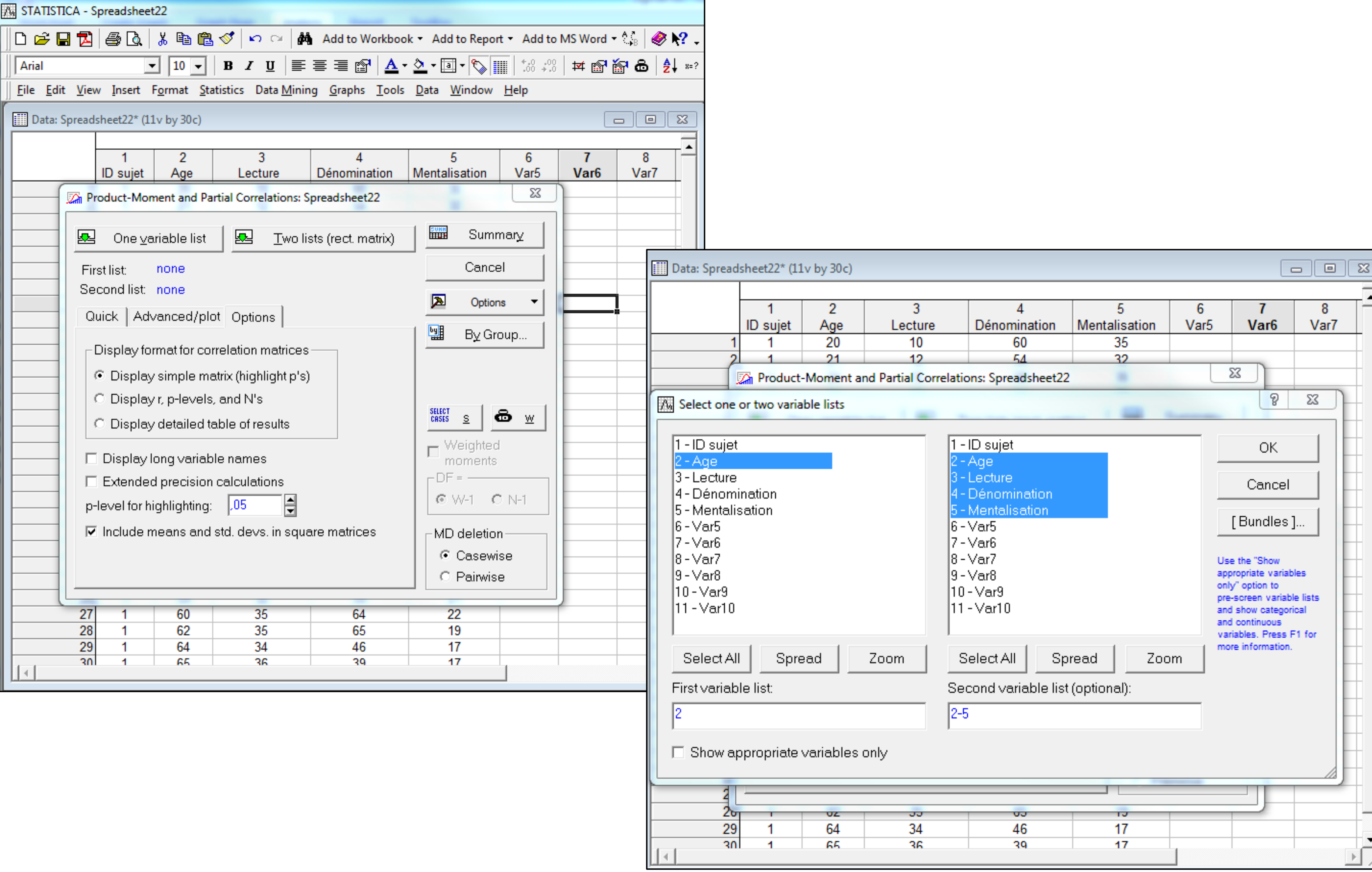
$$COV_{XY} = \frac{3177{,}40}{30-1} = 109{,}56$$

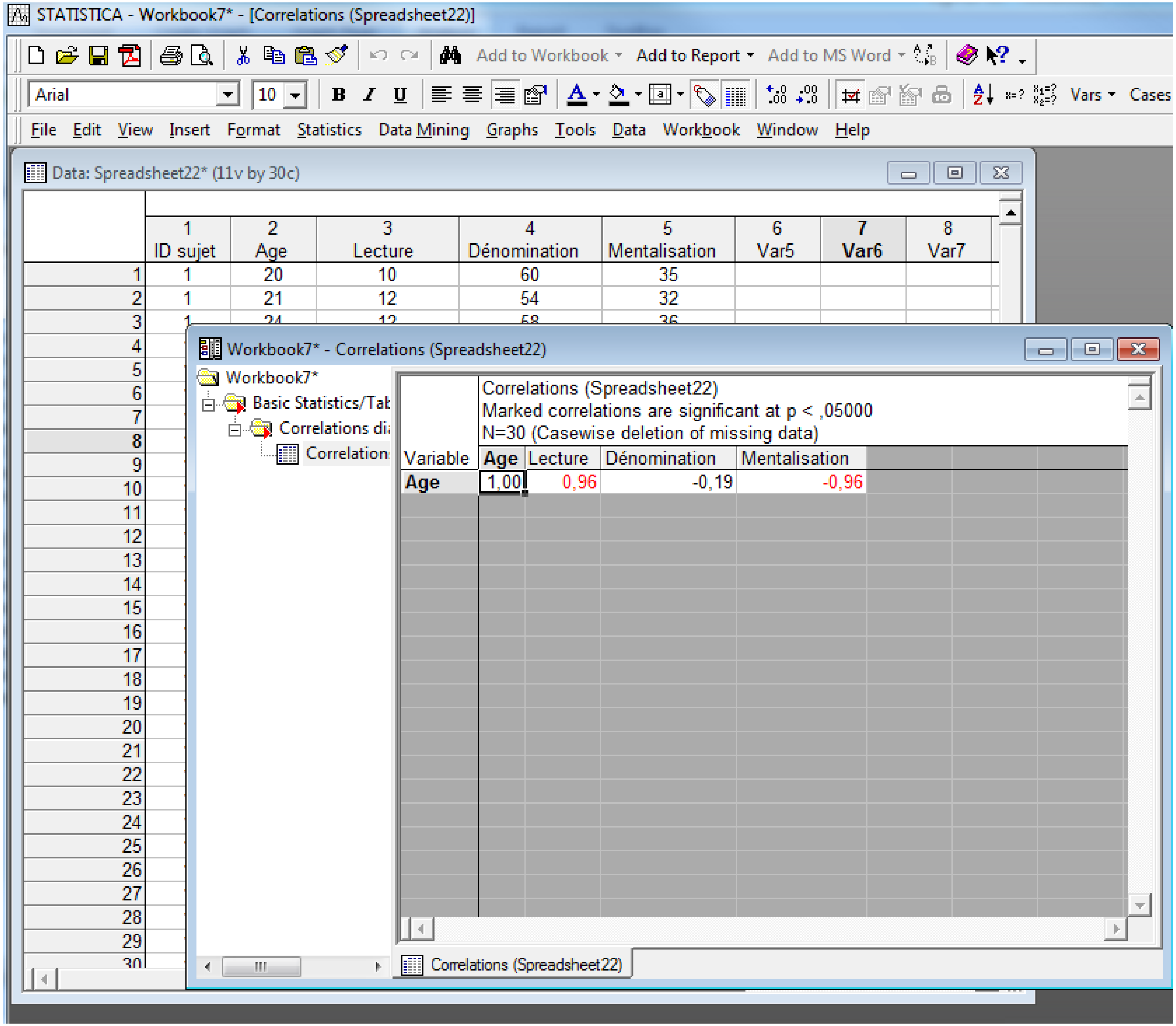
$$r = \frac{COV_{YX}}{S_X S_Y} = \frac{109{,}56}{13{,}79*8{,}28} = 0{,}96$$





	1 ID sujet	2 Age	3 Lecture	4 Dénomination	5 Mentalisation	6 Var5	7 Var6	8 Var7
1	1	20	10	60	35			
2	1	21	12	54	32			
3	1	24	12	58	36			
4	1	24	13	64	35			
5	1	26	13	50	34			
6	1	28	14	40	30			
7	1	30	11	70	32			
8	1	31	15	52	29			
9	1	32	17	50	26			
10	1	35	20	60	29			
11	1	36	16	65	30			
12	1	38	15	48	28			
13	1	40	20	47	27			
14	1	41	22	46	25			
15	1	42	21	53	26			
16	1	43	24	59	28			
17	1	46	25	62	24			
18	1	48	24	50	25			
19	1	49	18	34	24			
20	1	50	25	48	22			
21	1	51	26	49	24			
22	1	53	28	63	23			
23	1	55	30	52	22			
24	1	57	34	49	22			
25	1	58	29	48	22			
26	1	59	32	59	21			
27	1	60	35	64	22			
28	1	62	35	65	19			
29	1	64	34	46	17			
30	1	65	36	39	17			


STATISTICA - Spreadsheet22
Add to Workbook
Add to Report
Add to MS Word
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
Data: Spreadsheet22* (11v by 30c)
1 ID sujet
2 Age
3 Lecture
4 Dénomination
5 Mentalisation
6 Var5
7 Var6
8 Var7
Product-Moment and Partial Correlations: Spreadsheet22
One variable list
Two lists (rect. matrix)
Summary
Cancel
Options
By Group...
First list: none
Second list: none
Quick
Advanced/plot
Options
Display format for correlation matrices
Display simple matrix (highlight p's)
Display r, p-levels, and N's
Display detailed table of results
Display long variable names
Extended precision calculations
p-level for highlighting: ,05
Include means and std. devs. in square matrices
Weighted moments
DF =
W-1
N-1
MD deletion
Casewise
Pairwise
Select one or two variable lists
1 - ID sujet
2 - Age
3 - Lecture
4 - Dénomination
5 - Mentalisation
6 - Var5
7 - Var6
8 - Var7
9 - Var8
10 - Var9
11 - Var10
OK
Cancel
[Bundles]...
Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.
Select All
Spread
Zoom
First variable list:
2
Second variable list (optional):
2-5
Show appropriate variables only


STATISTICA - Workbook7* - [Correlations (Spreadsheet22)]
Add to Workbook
Add to Report
Add to MS Word
Arial
10
Vars
Cases
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Workbook Window Help
Data: Spreadsheet22* (11v by 30c)
1 ID sujet
2 Age
3 Lecture
4 Dénomination
5 Mentalisation
6 Var5
7 Var6
8 Var7
1 1 20 10 60 35
2 1 21 12 54 32
Workbook7* - Correlations (Spreadsheet22)
Workbook7*
Basic Statistics/Tab
Correlations di
Correlation
Correlations (Spreadsheet22)
Marked correlations are significant at p < ,05000
N=30 (Casewise deletion of missing data)
Variable
Age
Lecture
Dénomination
Mentalisation
Age 1,00 0,96 -0,19 -0,96
Correlations (Spreadsheet22)

6- Corrélation et régression linéaire

5-1 Le diagramme de dispersion

5-2 Le concept de covariance

6-3 Le coefficient de corrélation de Pearson

6-4 r ajusté

6-5 Droite de régression

6-6 la signification de r^2

6-7 Test de signification de r et b

6-8 Les facteurs influençant la corrélation

6- Corrélation et régression

❖ Le coefficient de corrélation que nous venons de calculer est celui qui est généralement utilisé. Toutefois, il s'agit *d'une estimation biaisée du coefficient de corrélation de la population* (puisque nous travaillons sur un échantillon).

❖ Nous devons donc **ajusté la corrélation**, avec la formule suivante:

$$r_{ajusté} = \sqrt{1 - \frac{(1-r^2)(N-1)}{N-2}} = \sqrt{1 - \frac{(1-0{,}96^2)(30-1)}{30-2}} = 0{,}96$$

❖ Si nous reprenons notre exemple de tout à l'heure :

$$r_{ajusté} = \sqrt{1 - \frac{(1-r^2)(N-1)}{N-2}} = \sqrt{1 - \frac{(1-0{,}96^2)(30-1)}{30-2}} = 0{,}96$$

Si *n* était tout petit et la corrélation moins forte:

$$r_{ajusté} = \sqrt{1 - \frac{(1-r^2)(N-1)}{N-2}} = \sqrt{1 - \frac{(1-\mathbf{0{,}55}^2)(\mathbf{10}-1)}{\mathbf{10}-2}} = 0{,}464$$

6- Corrélation et régression linéaire

5-1 Le diagramme de dispersion

5-2 Le concept de covariance

6-3 Le coefficient de corrélation de Pearson

6-4 r ajusté

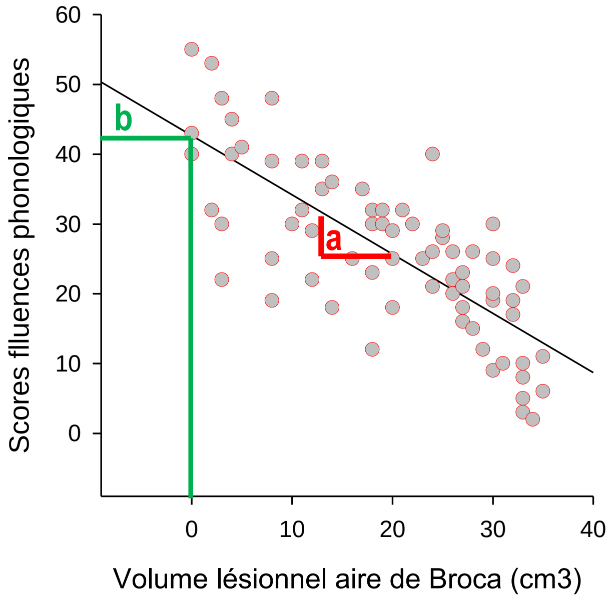
6-5 Droite de régression

6-6 la signification de r²

6-7 Test de signification de r et b

6-8 Les facteurs influençant la corrélation

6- Corrélation et régression



- Supposons que nous souhaitions prédire le score en fluences phonologiques de patients atteints d'une lésion de l'aire de Broca. Le graphique ci-contre représente les scores en fluences en fonction du volume lésionnel dans l'aire de Broca.

- La droite de régression linéaire va s'écrire : $\bar{Y} = \ bX + a$

  Avec $\bar{Y} =$ la valeur prédite de Y

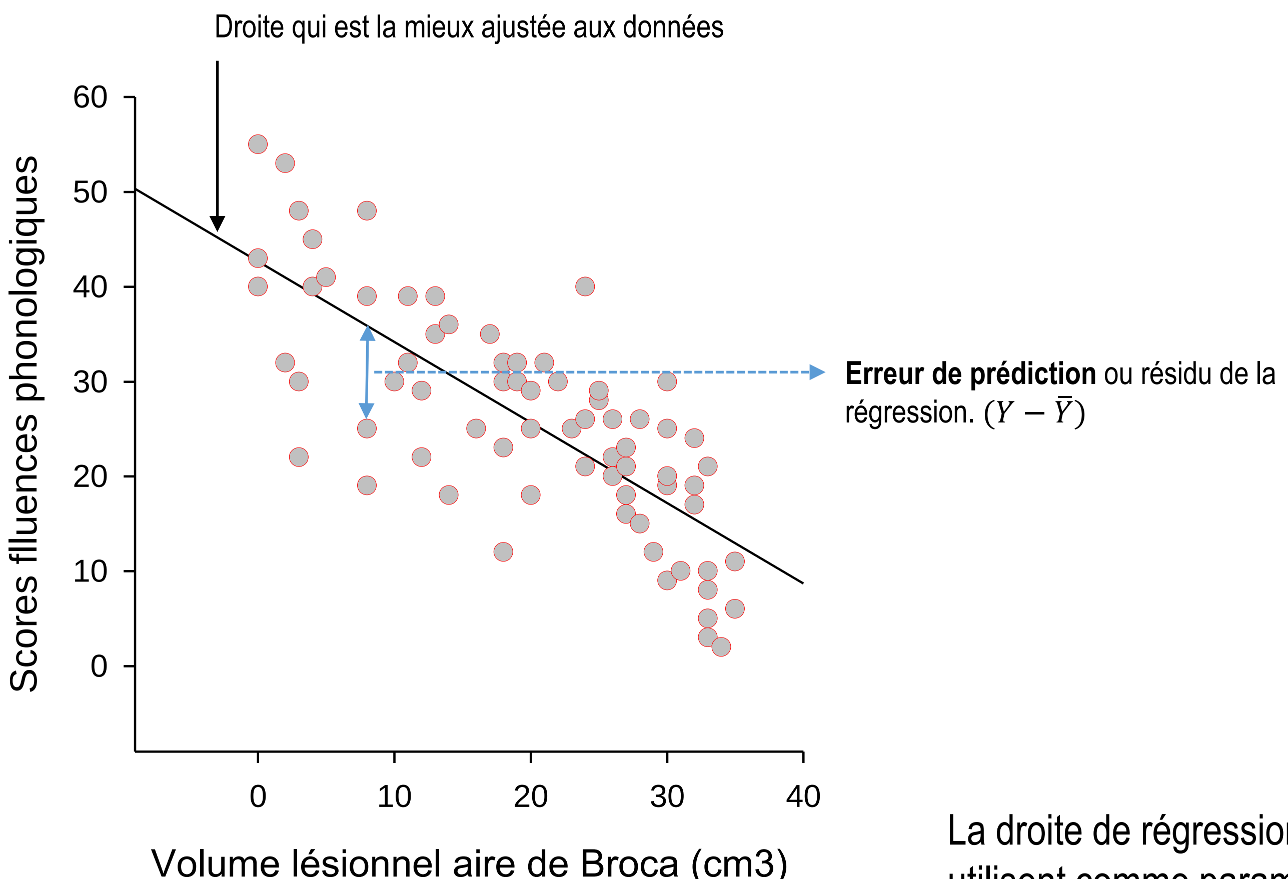
  b = **la pente** de la droite de régression

  a = **l'ordonnée à l'origine** (la valeur de $\bar{Y}$ lorsque X = 0)

  X = la valeur du prédicteur

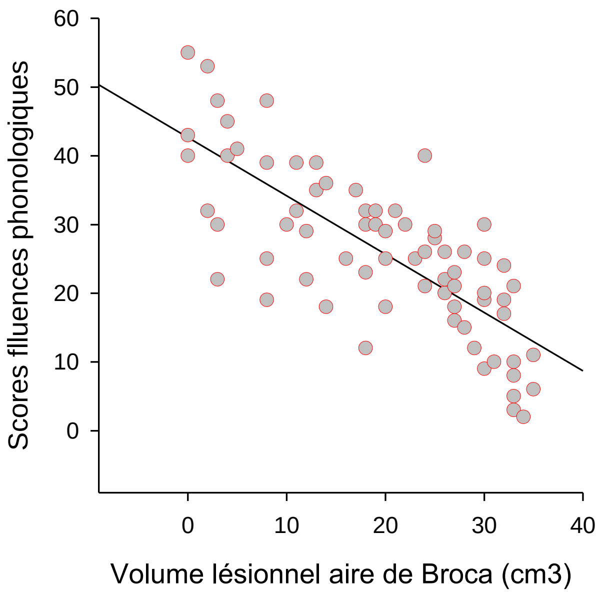
6.5 la droite de régression

6- Corrélation et régression



La droite de régression est la droite qui utilisent comme paramètres les valeurs optimales de *a* et de *b* en recherchant les valeurs de *a* et de *b* qui minimisent $\sum(Y - \bar{Y})^2$

6- Corrélation et régression



- Equations normales:

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

$$b = \frac{cov_{XY}}{s_X^2}$$

- Dans cet exemple :

$\bar{Y}$ = 19,65
$\bar{X}$ = 26
s_X = 10,62
cov_{xy}= -95,92

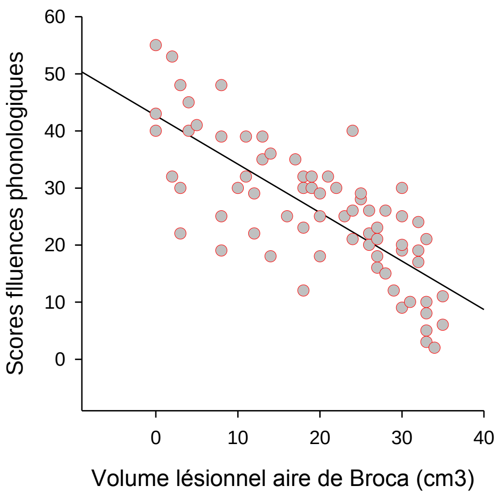
$$b = \frac{cov_{XY}}{s_X^2} = \frac{-95{,}92}{112{,}78} = -0{,}85$$

$a = \bar{Y} - b\bar{X}$ = 19,65 – (-0,85*26) = 41,75

$$\bar{Y} = -0{,}85X + 41{,}75$$

6.5 la droite de régression

6- Corrélation et régression



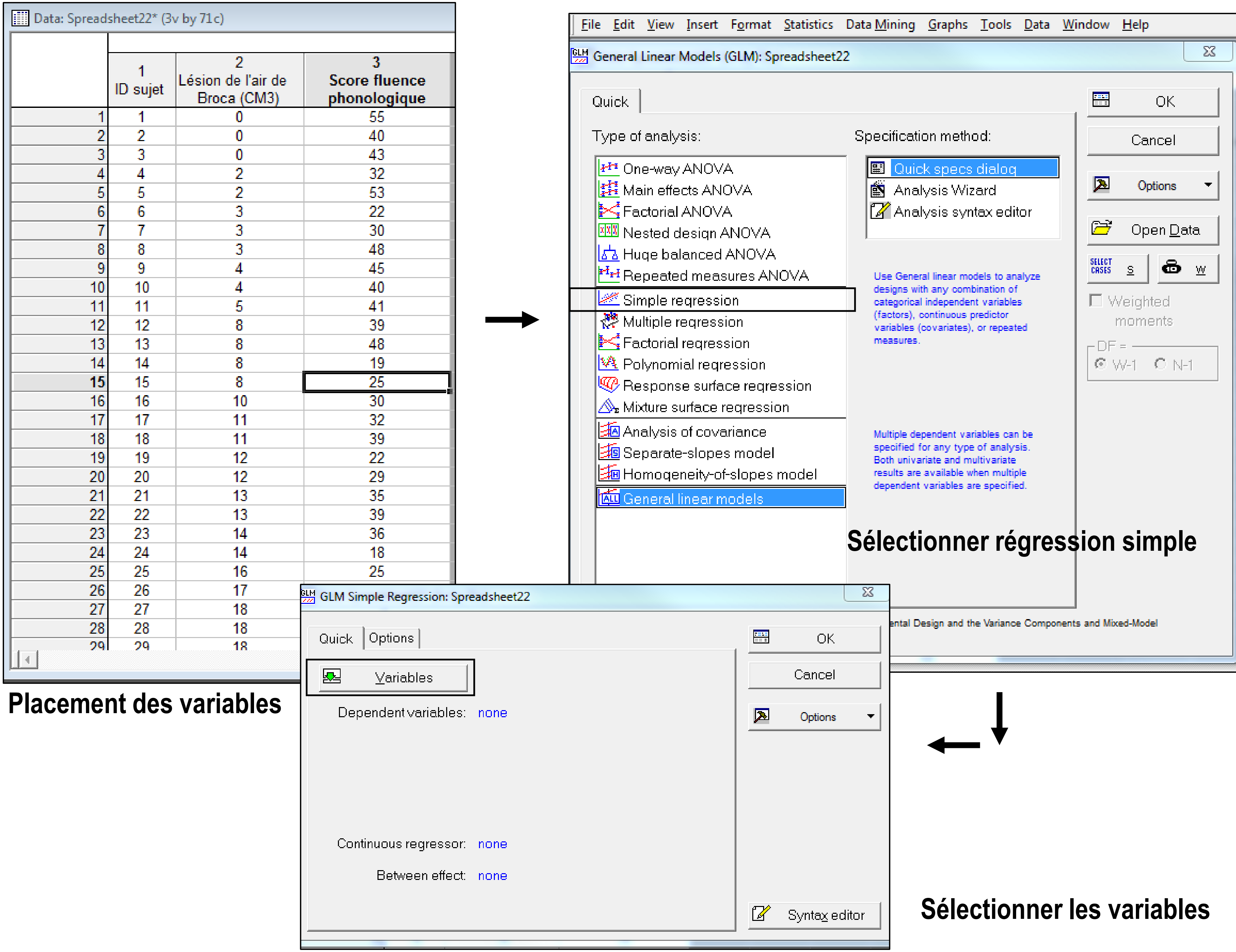
$$\bar{Y} = -0{,}85X + 41{,}75$$

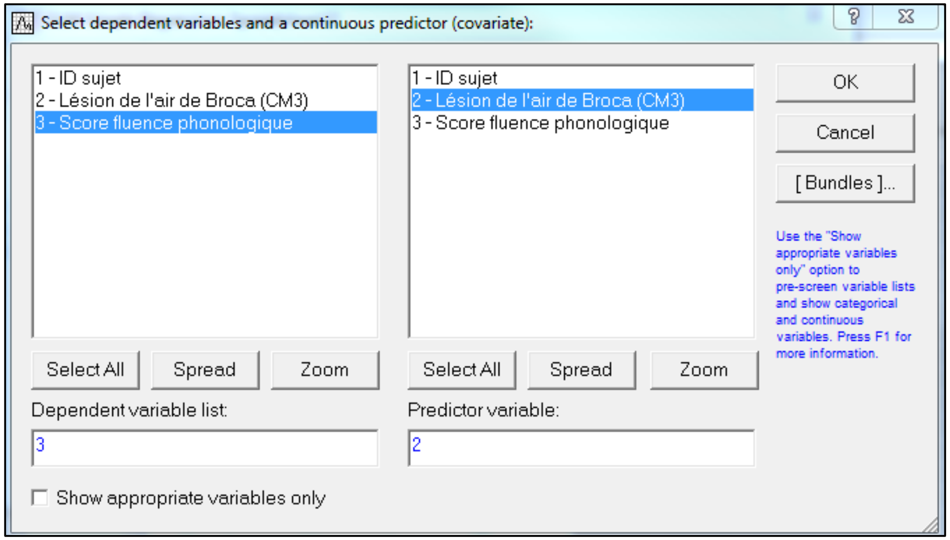
Pour X = 1 ; $\bar{Y}$ = 40,9
Pour X = 10 ; $\bar{Y}$ = 33,25
Pour X = 30 ; $\bar{Y}$ = 16,25
Pour X = 40 ; $\bar{Y}$ = 7,75

Remarque :

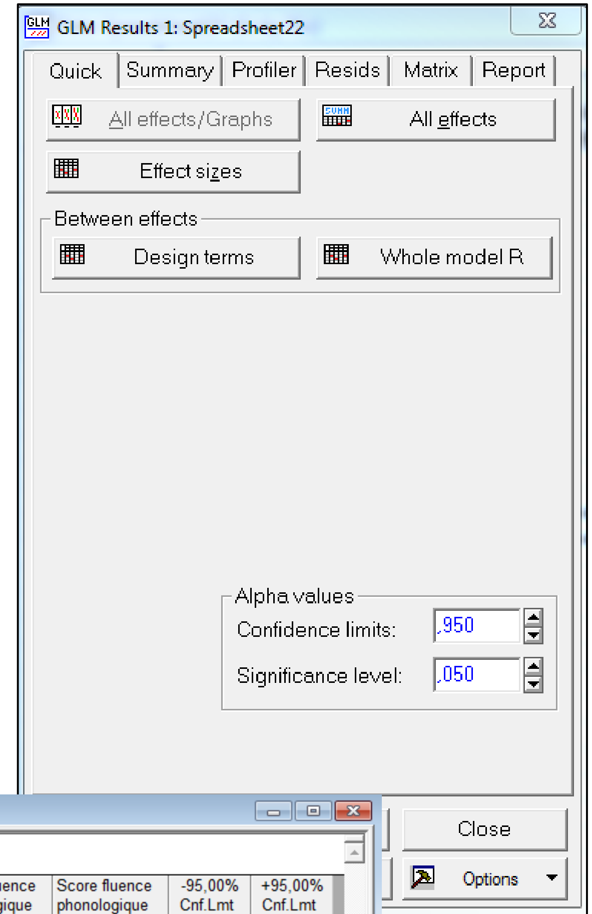
- La pente de la régression est une donnée particulièrement utile.

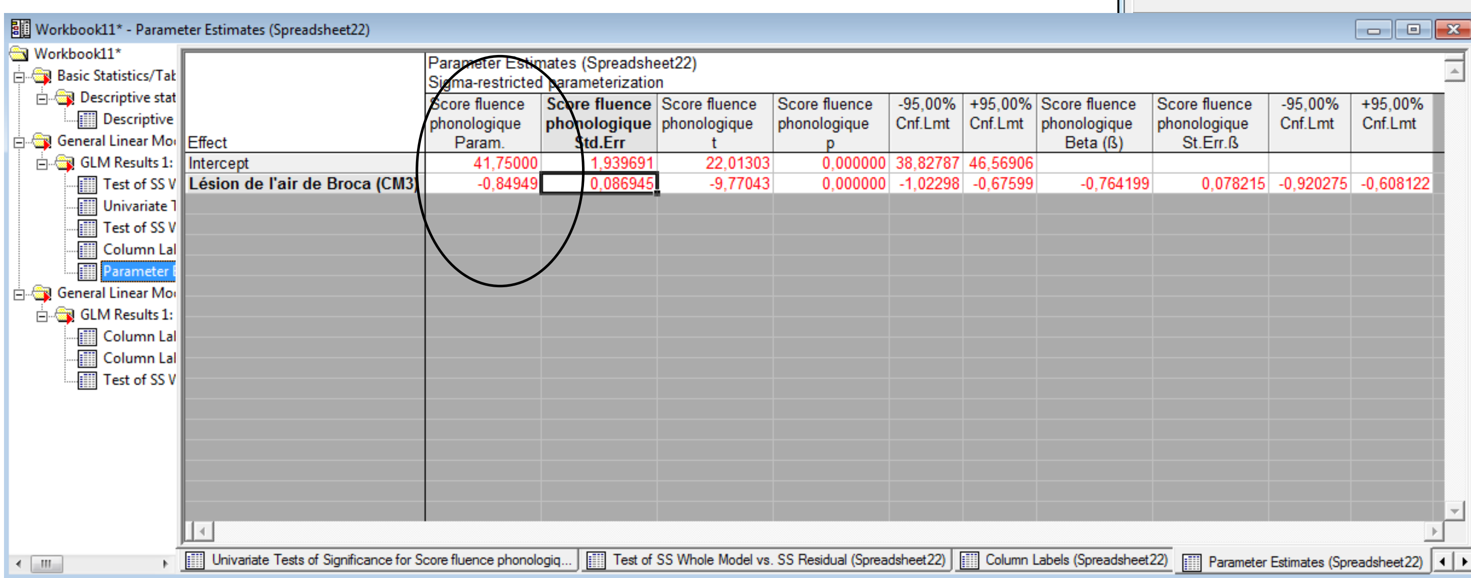
- Elle constitue une mesure du taux de changement prédit au niveau de Y. On sait ici que quand le volume lésionnelle dans l'aire de Broca augmente de 1 CM^3 le score en dénomination diminue de 0,85 point.


Data: Spreadsheet22* (3v by 71c)
1 ID sujet
2 Lésion de l'air de Broca (CM3)
3 Score fluence phonologique
1 1 0 55
2 2 0 40
3 3 0 43
4 4 2 32
5 5 2 53
6 6 3 22
7 7 3 30
8 8 3 48
9 9 4 45
10 10 4 40
11 11 5 41
12 12 8 39
13 13 8 48
14 14 8 19
15 15 8 25
16 16 10 30
17 17 11 32
18 18 11 39
19 19 12 22
20 20 12 29
21 21 13 35
22 22 13 39
23 23 14 36
24 24 14 18
25 25 16 25
26 26 17
27 27 18
28 28 18
29 29 18
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
General Linear Models (GLM): Spreadsheet22
Quick
Type of analysis:
One-way ANOVA
Main effects ANOVA
Factorial ANOVA
Nested design ANOVA
Huge balanced ANOVA
Repeated measures ANOVA
Simple regression
Multiple regression
Factorial regression
Polynomial regression
Response surface regression
Mixture surface regression
Analysis of covariance
Separate-slopes model
Homogeneity-of-slopes model
General linear models
Specification method:
Quick specs dialog
Analysis Wizard
Analysis syntax editor
Use General linear models to analyze designs with any combination of categorical independent variables (factors), continuous predictor variables (covariates), or repeated measures.
Multiple dependent variables can be specified for any type of analysis. Both univariate and multivariate results are available when multiple dependent variables are specified.
ental Design and the Variance Components and Mixed-Model
OK
Cancel
Options
Open Data
SELECT CASES S
W
Weighted moments
DF =
W-1
N-1
Sélectionner régression simple
GLM Simple Regression: Spreadsheet22
Quick
Options
Variables
Dependent variables: none
Continuous regressor: none
Between effect: none
OK
Cancel
Options
Syntax editor
Placement des variables
Sélectionner les variables



Sélectionner les variables: dépendante versus prédictive





6- Corrélation et régression linéaire

5-1 Le diagramme de dispersion

5-2 Le concept de covariance

6-3 Le coefficient de corrélation de Pearson

6-4 r ajusté

6-5 Droite de régression

6-6 la signification de r^2

6-7 Test de signification de r et b

6-8 Les facteurs influençant la corrélation

6- Corrélation et régression

- Calcul du *r²*

$$r^2 = \frac{SC_Y - SCr_{ésiduelle}}{SC_Y} \text{ ou } \frac{SC_{\bar{Y}}}{SC_Y}$$

Avec,

$SC_Y = \sum(Y - \bar{Y})^2$ *correspond à la variabilité des scores en dénomination*

$SC_{\hat{Y}} = \sum\left(\hat{Y} - \bar{Y}\right)^2$ *correspond à la variabilité des scores en dénomination imputable à la variabilité du volume lésionnel dans l'aire de Broca*

$SC_{résiduelle} = SC_Y - SC_{\hat{Y}}$ *correspond à la variabilité des scores en dénomination qui n'est pas imputable à la variabilité du volume lésionnel dans l'aire de Broca*

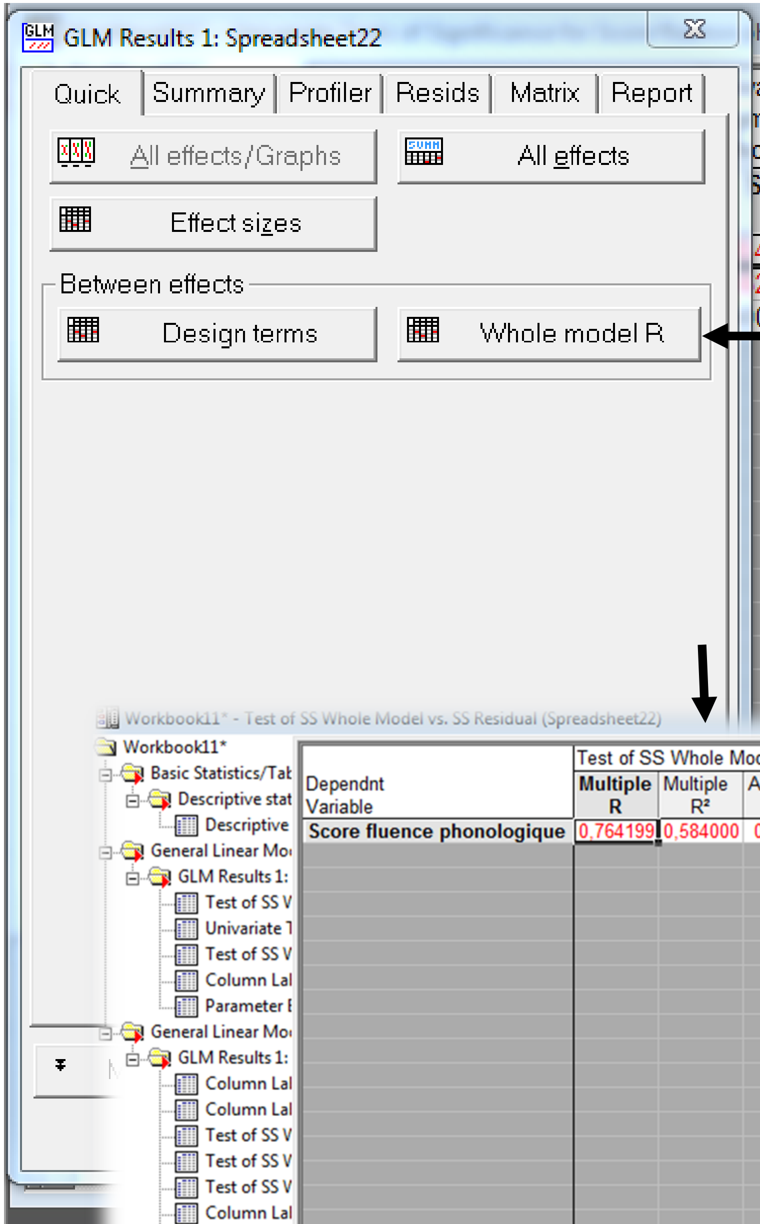
6- Corrélation et régression

6.6 Le r^2 comme mesure de la variabilité prédite

❖ Importance du r^2 *(dans notre exemple,* $r = 0.764$, $r^2 = 0,584$)

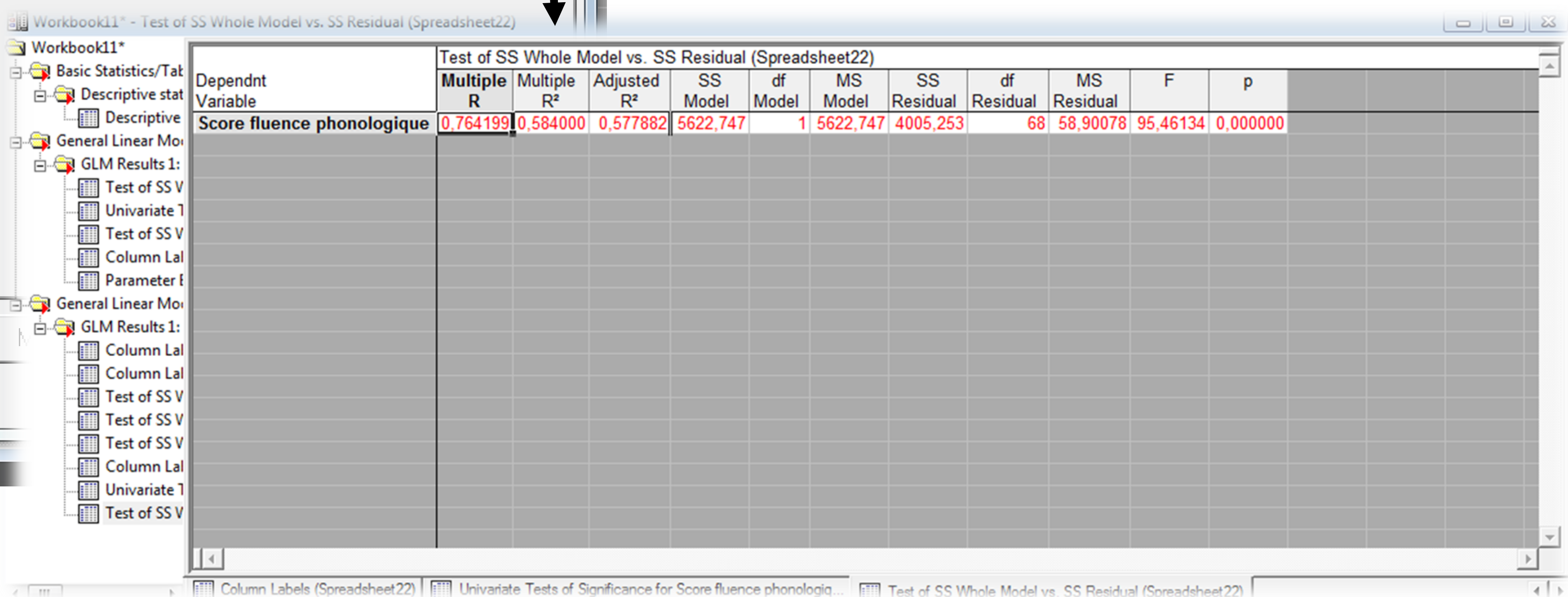
- Dans notre exemple, grâce au r^2, nous pouvons dire que $0,764^2 = 58,4\%$ de la variabilité dans les scores de dénomination peut être prédite grâce (donc est liée) au volume lésionnel dans l'aire de Broca, ce qui implique que 41,6% de la variabilité restante est imputable à d'autres facteurs. Notre modèle semble donc excellent.

- Si notre r était de 0,13, seulement $0,13^2 = 0,0169 = 1,69\%$ de variabilité des scores en dénomination serait expliqué par la variabilité de notre prédicteur. Dès lors, notre modèle serait inadéquate.



Sélectionner « *whole model R* »

Vous obtenez les valeurs de *r* et de *r²*



	Test of SS Whole Model vs. SS Residual (Spreadsheet22)										
Dependnt Variable	Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Score fluence phonologique	0,764199	0,584000	0,577882	5622,747	1	5622,747	4005,253	68	58,90078	95,46134	0,000000

6- Corrélation et régression linéaire

5-1 Le diagramme de dispersion

5-2 Le concept de covariance

6-3 Le coefficient de corrélation de Pearson

6-4 r ajusté

6-5 Droite de régression

6-6 la signification de r^2

6-7 Test de signification de r et b

6-8 Les facteurs influençant la corrélation

6- Corrélation et régression

- Dans le cadre de la régression linéaire, nous allons calculer la signification de *r* et de *b* de la même façon (ce qui n'est pas le cas si l'on se place dans le cadre des régressions multiples).

- Nous pouvons montrer que :

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Dans notre exemple,

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = 0{,}764\frac{\sqrt{70-2}}{\sqrt{1-0{,}764^2}} = 9{,}77 \ (p < 0{,}00000001)$$

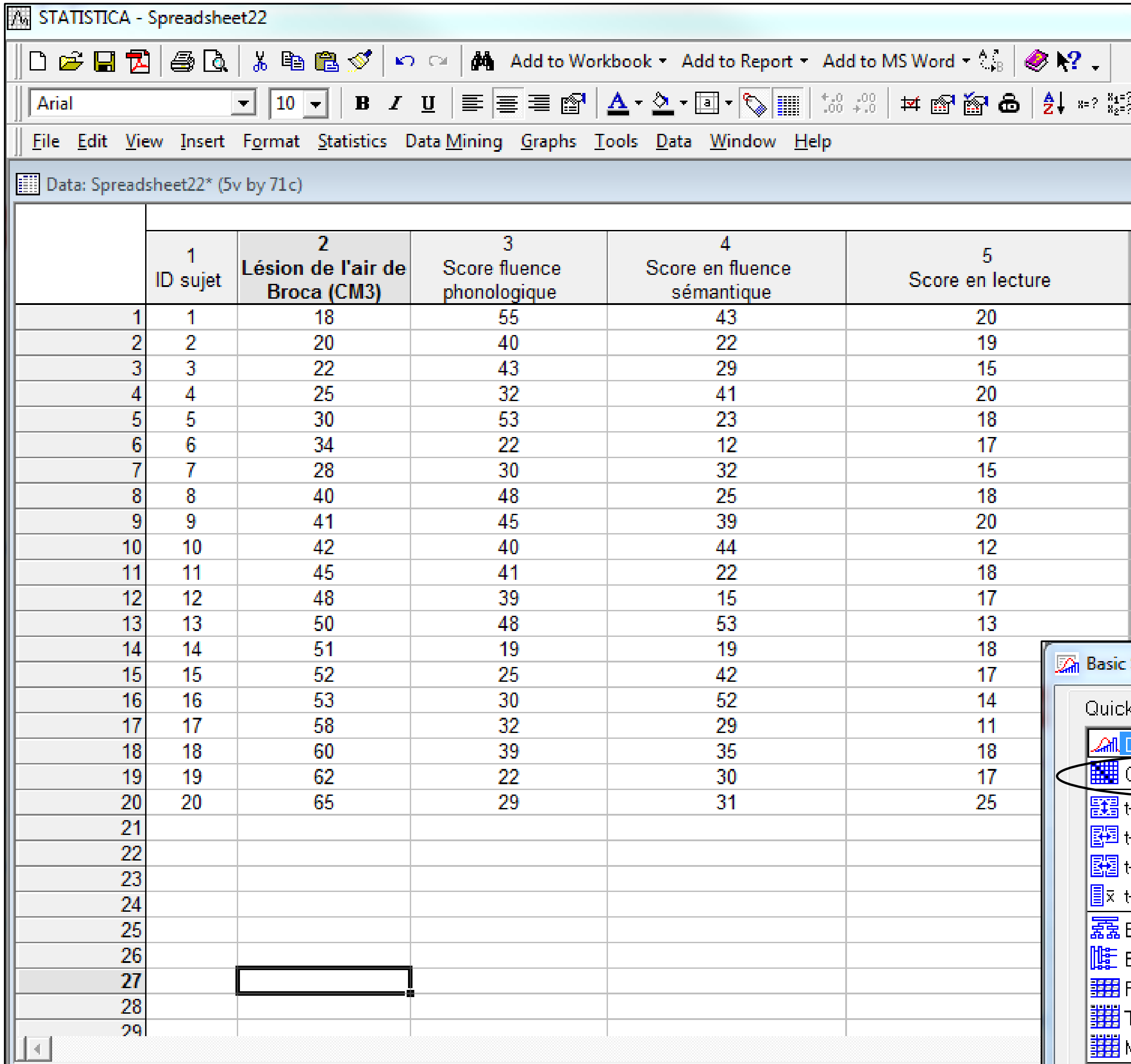
Workbook11* - Parameter Estimates (Spreadsheet22)

Workbook11*
- Basic Statistics/Tab
  - Descriptive stat
    - Descriptive
- General Linear Mo
  - GLM Results 1:
    - Test of SS W
    - Univariate T
    - Test of SS W
    - Column Lal
    - Parameter E
- General Linear Mo
  - GLM Results 1:
    - Column Lal
    - Column Lal
    - Test of SS W

Parameter Estimates (Spreadsheet22)
Sigma-restricted parameterization

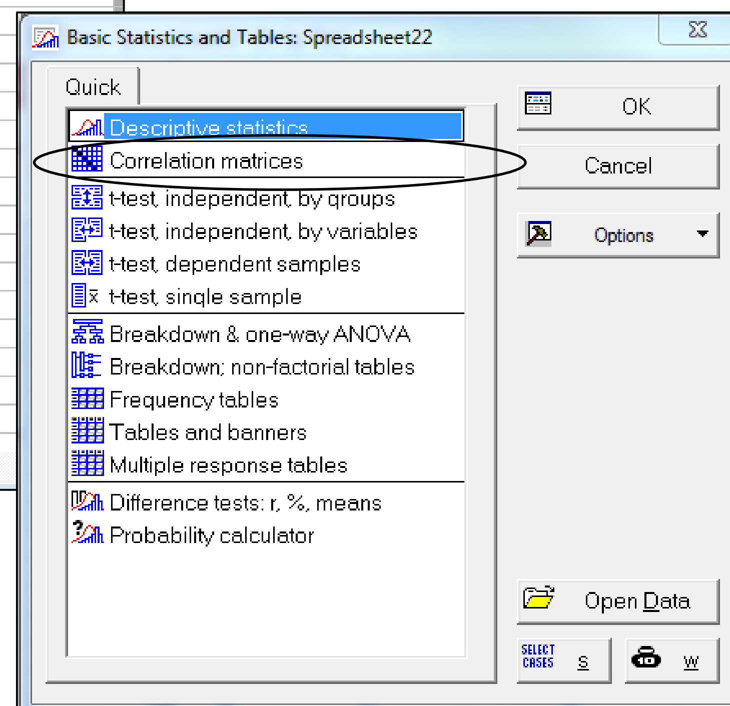
Effect	Score fluence phonologique Param.	Score fluence phonologique Std.Err	Score fluence phonologique t	Score fluence phonologique p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt	Score fluence phonologique Beta (ß)	Score fluence phonologique St.Err.ß	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Intercept	41,75000	1,939691	22,01303	0,000000	38,82787	46,56906				
Lésion de l'air de Broca (CM3)	-0,84949	0,086945	-9,77043	0,000000	-1,02298	-0,67599	-0,764199	0,078215	-0,920275	-0,608122

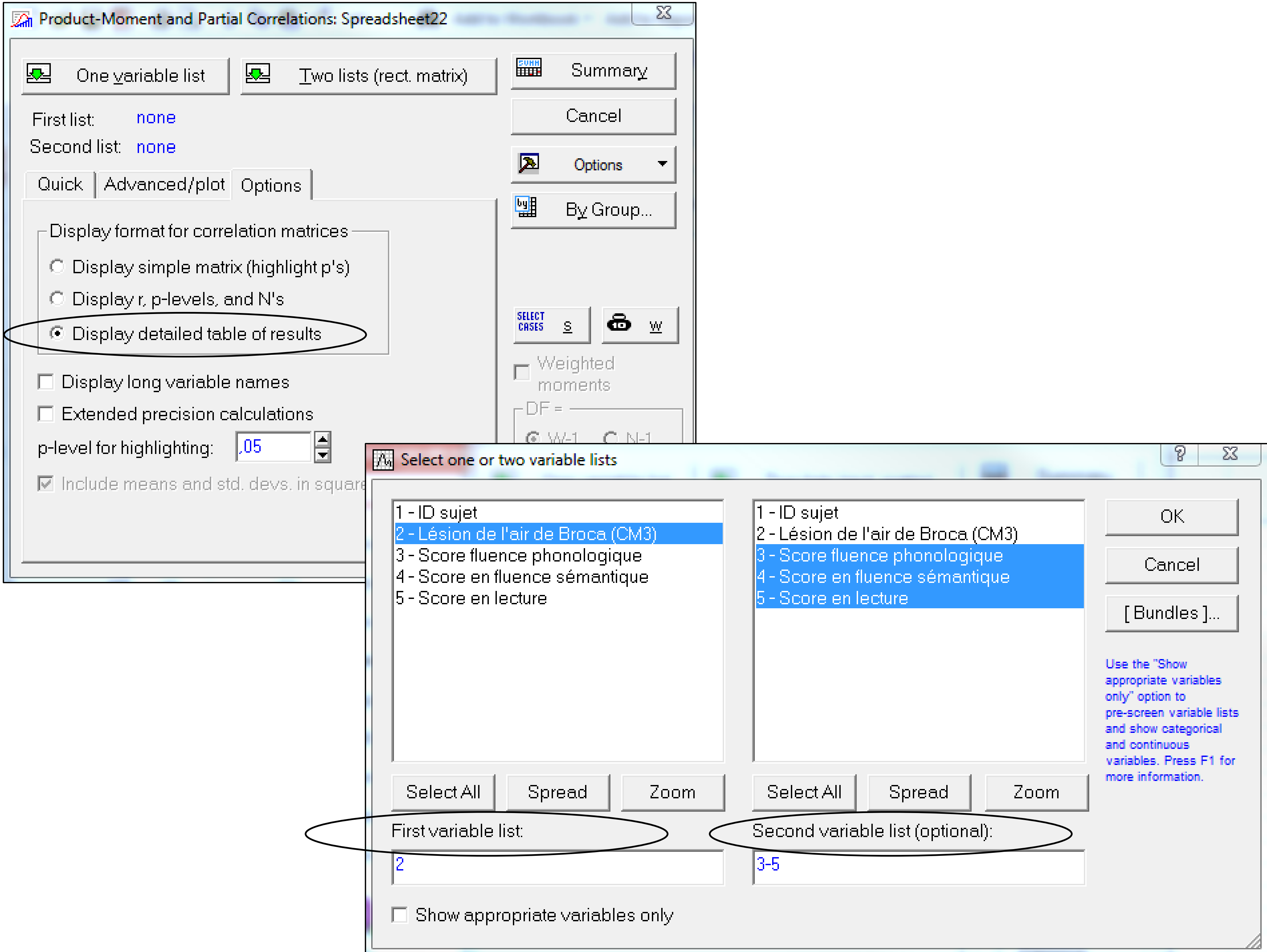
Univariate Tests of Significance for Score fluence phonologiq... | Test of SS Whole Model vs. SS Residual (Spreadsheet22) | Column Labels (Spreadsheet22) | Parameter Estimates (Spreadsheet22)



	1 ID sujet	2 Lésion de l'air de Broca (CM3)	3 Score fluence phonologique	4 Score en fluence sémantique	5 Score en lecture
1	1	18	55	43	20
2	2	20	40	22	19
3	3	22	43	29	15
4	4	25	32	41	20
5	5	30	53	23	18
6	6	34	22	12	17
7	7	28	30	32	15
8	8	40	48	25	18
9	9	41	45	39	20
10	10	42	40	44	12
11	11	45	41	22	18
12	12	48	39	15	17
13	13	50	48	53	13
14	14	51	19	19	18
15	15	52	25	42	17
16	16	53	30	52	14
17	17	58	32	29	11
18	18	60	39	35	18
19	19	62	22	30	17
20	20	65	29	31	25
21					
22					
23					
24					
25					
26					
27					
28					
29					

Nouvelles données




Product-Moment and Partial Correlations: Spreadsheet22
One variable list
Two lists (rect. matrix)
Summary
Cancel
Options
By Group...
First list: none
Second list: none
Quick
Advanced/plot
Options
Display format for correlation matrices
Display simple matrix (highlight p's)
Display r, p-levels, and N's
Display detailed table of results
Display long variable names
Extended precision calculations
p-level for highlighting: ,05
Include means and std. devs. in square
SELECT CASES
S
W
Weighted moments
DF =
Select one or two variable lists
1 - ID sujet
2 - Lésion de l'air de Broca (CM3)
3 - Score fluence phonologique
4 - Score en fluence sémantique
5 - Score en lecture
1 - ID sujet
2 - Lésion de l'air de Broca (CM3)
3 - Score fluence phonologique
4 - Score en fluence sémantique
5 - Score en lecture
OK
Cancel
[Bundles]...
Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.
Select All
Spread
Zoom
Select All
Spread
Zoom
First variable list:
2
Second variable list (optional):
3-5
Show appropriate variables only

Qu'en déduisez-vous ?

rrelations (Spreadsheet22)

Correlations (Spreadsheet22)
Marked correlations are significant at $p < ,05000$
(Casewise deletion of missing data)

Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	r²	t	p	N	Constant dep: Y	Slope dep: Y	Constant dep: X	Slope dep: X
Lésion de l'air de Broca (CM3)	42,20000	14,64887									
Score fluence phonologique	36,60000	10,49010	-0,448814	0,201434	-2,13082	0,047147	20	50,16295	-0,321397	65,13889	-0,626746
Lésion de l'air de Broca (CM3)	42,20000	14,64887									
Score en fluence sémantique	31,90000	11,59809	0,099254	0,009851	0,42319	0,677171	20	28,58378	0,078583	38,20095	0,125362
Lésion de l'air de Broca (CM3)	42,20000	14,64887									
Score en lecture	17,10000	3,19374	-0,071323	0,005087	-0,30337	0,765085	20	17,75621	-0,015550	47,79412	-0,327141

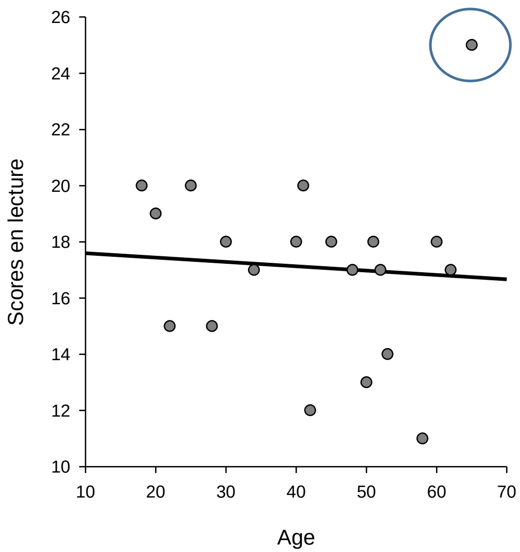
6- Corrélation et régression linéaire

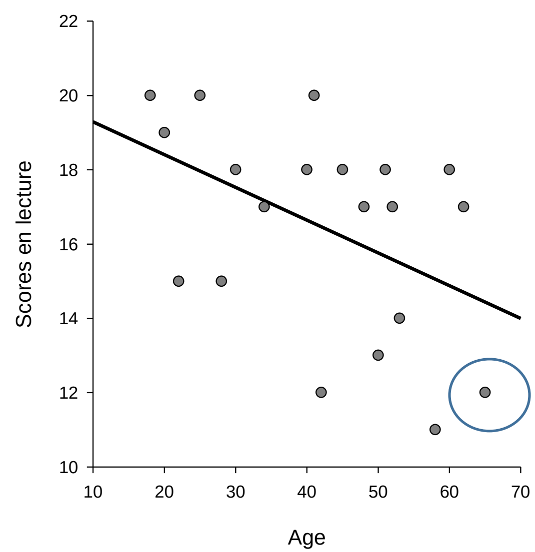
6-1 Le diagramme de dispersion
6-2 Le concept de covariance
6-3 Le coefficient de corrélation de Pearson
6-4 r ajusté
6-5 Droite de régression
6-6 la signification de r^2
6-7 Test de signification de r et b
6-8 Les facteurs influençant la corrélation

6- Corrélation et régression

❖ Les valeurs extrêmes.

Les scores extrêmes sur une distribution influencent considérablement les résultats.





***r* = 0,0713, *r²* = 0,00509, *t* = -0,303, *p* = 0,765**

***r* = 0,462, *r²* = 0,213, *t* = -2,208, *p* = 0,040**

Le modèle est non-significatif

Le modèle est significatif

7- Comparer plusieurs groupes : Analyse de variance