

# **Relation entre une variable continue et une variable catégorielle**

Pierre-Yves Henry  
henry@mnhn.fr



# Quelle analyse ?

Y \ X	Var. catégorielle	Var. continue
	Var. catégorielle	Var. continue
Var. catégorielle	Analyse de fréquence	Faire des catégories  Transformer en variable quantitative
Var. continue	<b>ANOVA</b> <b>Comparaison de moyennes</b>	Corrélation Régression

# L'ANOVA

© Julien GONIN



**(ANalysis Of VAriance)**

# ANOVA

## Test de la relation entre une variable continue et une variable catégorielle

Ex: relation entre indice de condition corporelle (Y) et âge (X) – données ERIRUB

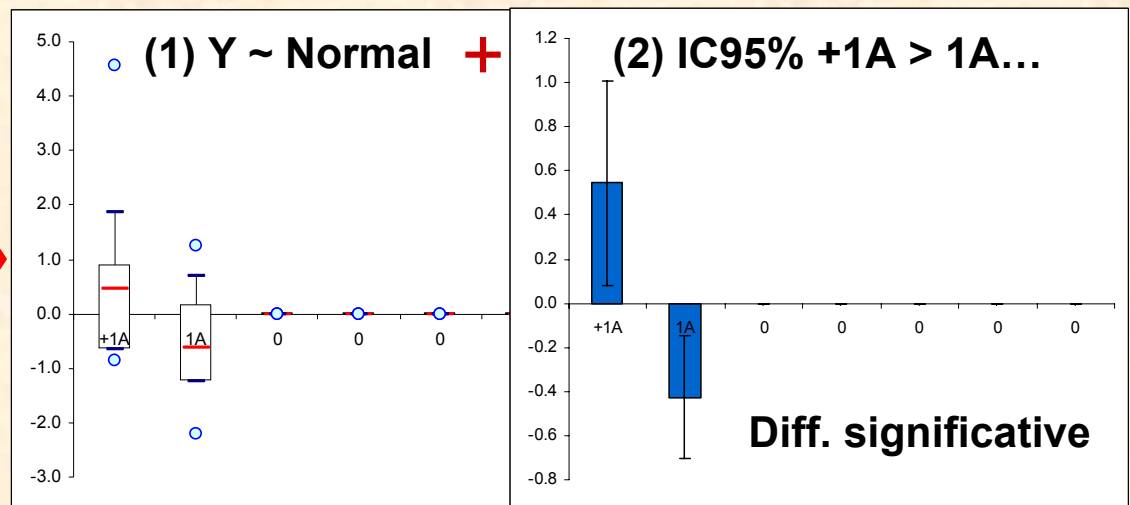
Conditions d'application:  
Y ~ Normal par catégorie  
Variance  $\approx$  entre catégories

Etape 1: visualisation graphique de la relation entre les deux variables et évaluation des conditions d'application

Utiliser le fichier "2 2a Statistiques descriptives et exploration graphique.xls"

Coller les données dans la page "Données"

	A	B	C	D	E	F	G	H
1	Groupe	1	2					
2	Nom de groupe	+1A	1A					
3		-0.14	0.85					
4		0.45	-0.85					
5		1.00	0.40					
6		0.75	0.50					
7		0.00	-0.70					
8		2.05	-1.05					
9		0.20	-0.60					
10		0.01	-1.40					
11		0.95	-0.70					
12		-0.70	-1.10					
13		0.61	1.26					
14		2.45	-0.45					
15		0.15	-0.70					
16		0.90	-1.25					
17		0.70	-1.15					



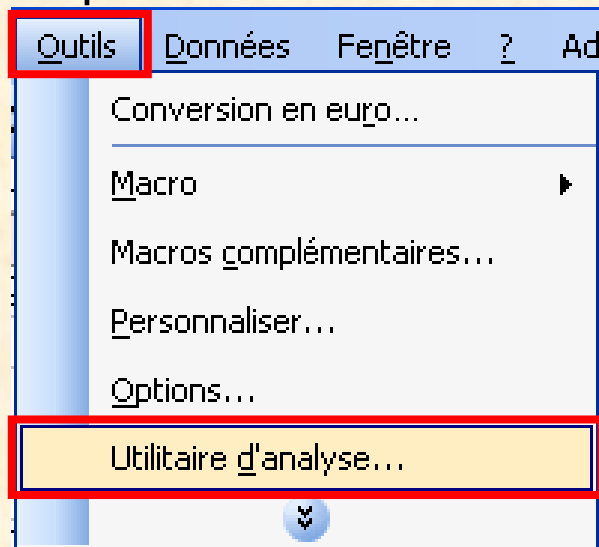
# ANOVA

Etape 2: préparer les données au bon format pour l'Utilitaire d'analyse

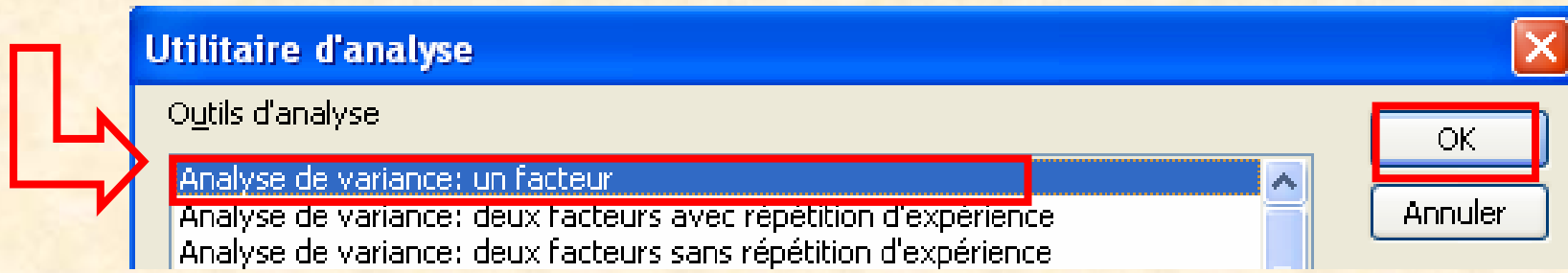
Utiliser le fichier "2 4 test ANOVA.xls"

Importer les données dans la page "données en forme pour analyse"

Etape 3: Ouvrir l'Utilitaire d'analyse



et choisir  
"Analyse de variance: un facteur"



	A	B
1	1A	+1A
2	0.85	-0.14
3	-0.85	0.45
4	0.40	1.00
5	0.50	0.75
6	-0.70	0.00
7	-1.05	2.05
8	-0.60	0.20
9	-1.40	0.01
10	-0.70	0.95
11	-1.10	-0.70
12	1.26	0.61
13	-0.45	2.45
14	-0.70	0.15
15	-1.25	0.90
16	-1.15	0.70
17	0.15	-0.39
18	0.50	-0.65
19	-0.64	0.66
20	0.20	0.05
21	-0.85	1.55
22	-2.20	-0.85
23	-0.10	0.51
24	-0.64	-0.65
25	0.75	4.55
26	-0.95	-0.54
27	0.10	
28	-1.75	
29	-0.40	
30	-0.35	

# ANOVA

## Etape 4: définir la présentation des données

	A	B
1	1A	+1A
2	0.85	-0.14
3	-0.85	0.45
4	0.40	1.00
5	0.50	0.75
6	-0.70	0.00
7	-1.05	2.05
8	-0.60	0.20
9	-1.40	0.01
10	-0.70	0.95
11	-1.10	-0.70
12	1.26	0.61
13	-0.45	2.45
14	-0.70	0.15
15	-1.25	0.90
16	-1.15	0.70
17	0.15	-0.39
18	0.50	-0.65
19	-0.64	0.66
20	0.20	0.05
21	-0.85	1.55
22	-2.20	-0.85
23	-0.10	0.51
24	-0.64	-0.65
25	0.75	4.55
26	-0.95	-0.54
27	0.10	
28	-1.75	
29	-0.40	
30	-0.35	

**Analyse de variance: un facteur**

**Paramètres d'entrée**

Plage d'entrée:

Groupées par:

☒ Colonnes

☐ Lignes

☒ Intitulés en première ligne

Seuil de signification:

**Options de sortie**

☐ Plage de sortie:

☒ Insérer une nouvelle feuille:

☐ Créer un nouveau classeur

OK Annuler Aide



# ANOVA

## Etape 5: interprétation des résultats de l'ANOVA

### (1) Moyenne et variance pour chacun des groupes

#### RAPPORT DÉTAILLÉ

<i><b>Groupes</b></i>	<i><b>Nb d'éch</b></i>	<i><b>Somme</b></i>	<i><b>Moyenne</b></i>	<i><b>Variance</b></i>
<b>1A</b>	32	-13.661	<b>-0.427</b>	<b>0.654</b>
<b>+1A</b>	25	13.661	<b>0.546</b>	<b>1.398</b>
Total	57			

*Attention: la variance est 2 fois plus grand pour les +1A que pour les 1A*

*Le risque = ne pas trouver de différence significative alors qu'il y en aurait une*

# ANOVA

## Etape 5: interprétation des résultats de l'ANOVA

Principe: déterminer si la part de variation de Y expliquée par les variations de X est significative (c'est à dire importante par rapport au reste de variation à expliquer, soit non due au hasard)

Variation expliquée par  
X (l'âge)

(2) Test de l'effet de X (âge) sur Y (l'indice de condition corporelle)

ANALYSE DE VARIANCE

Source des variations	mmé des car	Nb ddl	ymé des car	F	Prob	Valeur critique pour F
Entre Groupes (= effet de l'âge)	13.297	1	13.297	13.584	0.001	4.016
A l'intérieur des groupes	53.839	55	0.979			
Total	67.136	56				

Variation **non** expliquée  
par X (l'âge)



# ANOVA

## Etape 5: interprétation des résultats de l'ANOVA

$H_0$ : la distribution de Y est la même pour toutes les catégories de X (la variation de Y entre catégories de X est mineure et peut-être expliquée par du hasard)

**Statistique F** = rapport des parts de variation expliquée / non expliquée

(2) Test de l'effet de X (âge) sur Y (l'indice de condition corporelle)

ANALYSE DE VARIANCE

Source des variations	mmé des car	Nb ddl	enne des car	F	Prob	Valeur critique pour F
Entre Groupes (= effet de l'âge)	13.297	1	13.297	13.584	0.001	4.016
A l'intérieur des groupes	53.839	55	0.979			
Total	67.136	56				

**Nb ddl**

**Probabilité que  $H_0$  soit vraie**

(1) effet de X = nb de catégories – 1

(2) variation résiduelle = N – ddl(X) - 1

# ANOVA

## Etape 5: interprétation des résultats de l'ANOVA

### Notations:

L'indice de condition corporelle est significativement différent entre juvéniles et adultes ( $F_{1,55} = 13.584$ ,  $P = 0.001$ )

L'indice de condition corporelle varie avec l'âge ( $F_{1,55} = 13.584$ ,  $P = 0.001$ )

chez les rougegorges familiers lors de la période de reproduction 2007-2008 sur la station STOC n°

### (2) Test de l'effet de X (âge) sur Y (l'indice de condition corporelle)

#### ANALYSE DE VARIANCE

Source des variations	mmé des car	Nb ddl	enne des car	F	Prob	Valeur critique pour F
Entre Groupes (= effet de l'âge)	13.297	1	13.297	13.584	0.001	4.016
A l'intérieur des groupes	53.839	55	0.979			
Total	67.136	56				

# Bonus 1

## Variances différentes

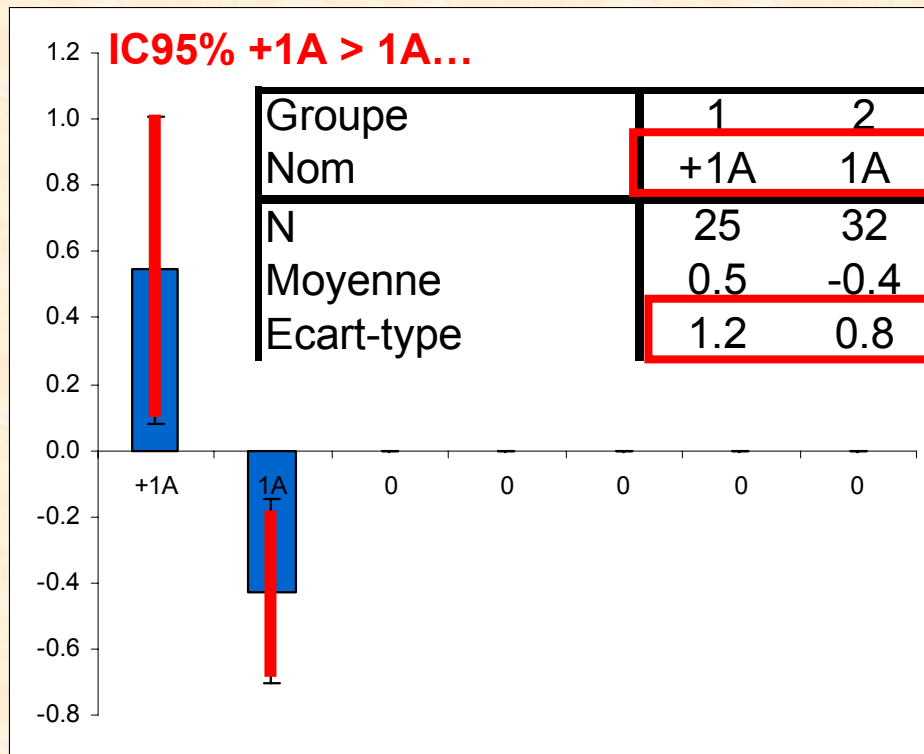


# 1- Test d'homogénéité de variance

## Test de la relation entre une variable continue et une variable catégorielle par l'ANOVA

Conditions d'application ANOVA:  
Y ~ Normal par catégorie  
Variance  $\approx$  entre catégories

Etape 1: visualisation graphique de la relation entre les deux variables et évaluation des conditions d'application



## Test d'égalité des variances

### Utilitaire d'analyse

#### Outils d'analyse

- Analyse de variance: deux facteurs sans
- Analyse de corrélation
- Analyse de covariance
- Statistiques descriptives
- Lissage exponentiel
- Test d'égalité des variances (F-Test)
- Transformation de Fourier Rapide (FFT)
- Histogramme
- Moyenne mobile
- Génération de nombres aléatoires

# 2- Test de moyenne avec variances différentes

Etape 2: interprétation du test d'égalité des variances

Test d'égalité des variances (F-Test)

	+1A	1A
Moyenne	0.54645115	-0.42691496
Variance	1.39829024	0.65419303
Observations	25	32
Degré de liberté	24	31
F	2.1374276	
P(F<=f) unilatéral	0.02372767	
Valeur critique pour F (	1.87507256	

La variance de l'indice de condition corporelle varie avec l'âge ( $F_{24,31} = 2.137$ ,  $P = 0.024$ )

## Utilitaire d'analyse

### Outils d'analyse

- Histogramme
- Moyenne mobile
- Génération de nombres aléatoires
- Analyse de position
- Régression linéaire
- Échantillonnage
- Test d'égalité des espérances: observations paires
- Test d'égalité des espérances: deux observations de variances égales
- Test d'égalité des espérances: deux observations de variances différentes
- Test de la différence significative minimale (z-test)

## 2- Test de moyenne avec variances différentes

Etape 3: définition des données ET de l'analyse

Test d'égalité des espérances: deux observations de variances dif...

Paramètres d'entrée

Plage pour la variable 1: \$B\$2:\$B\$27

Plage pour la variable 2: \$C\$2:\$C\$34

Différence entre les moyennes (hypothèse): 0

☒ Intitulé présent

Seuil de signification: 0.05

OK

Annuler

Aide

Définition de l' $H_0$  qui correspond à la question

0  $\leftrightarrow$  pas de différence

2  $\leftrightarrow$  différence

normalement attendue (p. ex. d'après autres études)

# 2- Test de moyenne avec variances différentes

## Etape 4: interprétation des résultats

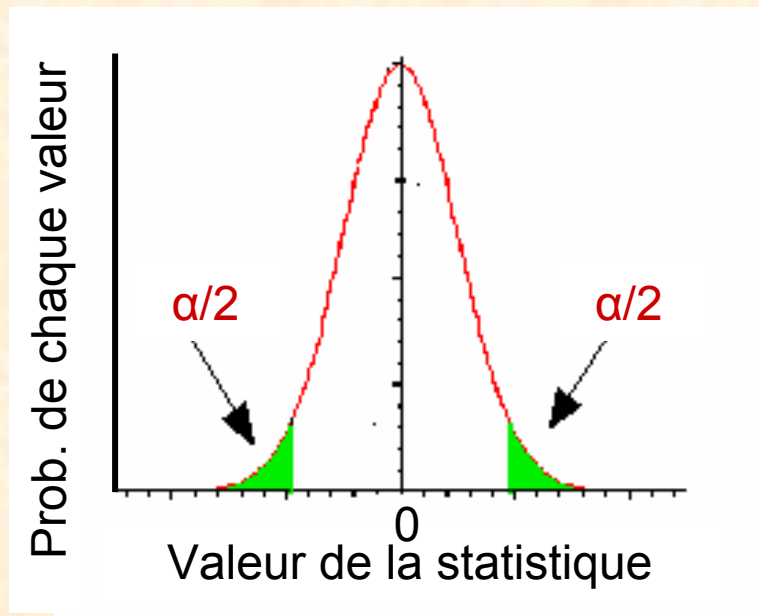
Test d'égalité des espérances: deux observations de variances différentes

	+1A	1A
Moyenne	0.54645115	-0.42691496
Variance	1.39829024	0.65419303
Observations	25	32
Différence hypothétique des m	0	
Degré de liberté	41	
Statistique t	3.52208807	
P(T<=t) unilatéral	0.000533	
Valeur critique de t (unilatéral)	1.682878	
P(T<=t) bilatéral	0.00106601	
Valeur critique de t (bilatéral)	2.01954095	

# 3- Test bilatéral ou unilatéral ?

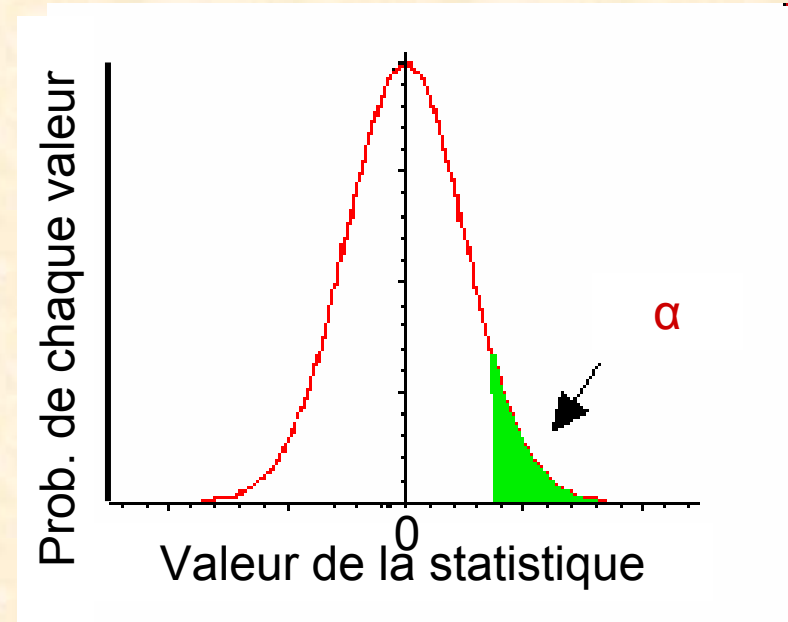
**Différence entre test bilatéral et test unilatéral:  
choix de deux hypothèses alternatives différentes**

Test **BILATERAL**



$H_1$  = différence entre  
groupes 1 et 2

Test **UNILATERAL**



$H_1$  = la moyenne du groupe  
2 est supérieure à celle du  
groupe 1

**Le choix se fait en fonction des connaissances biologiques  
de l'attendu**

**Pour l'âge: unilatéral OK car on s'attend à une croissance de l'aile pliée**

## 2- Test de moyenne avec variances différentes

### Etape 4: interprétation des résultats

Test d'égalité des espérances: deux observations de variances différentes

	+1A	1A
Moyenne	0.54645115	-0.42691496
Variance	1.39829024	0.65419303
Observations	25	32
Différence hypothétique des m	0	
Degré de liberté	41	
Statistique t	3.52208807	
P(T<=t) unilatéral	0.000533	
Valeur critique de t (unilatéral)	1.682878	
P(T<=t) bilatéral	0.00106601	
Valeur critique de t (bilatéral)	2.01954095	

# Bonus 2

**Quand les échantillons ne  
sont pas indépendants**

Photo : Thierry BARA

# Test de moyenne pour éch. pairés

## Test de la relation entre une variable continue (avec mesures non-indépendantes) et une variable catégorielle

**Cas classiques:**  
mesures sur les mêmes individus

**Conditions d'application ANOVA:**

$Y \sim \text{Normal}$  par catégorie

Variance  $\approx$  entre catégories

**Mesures indépendantes**

**Ex: test de l'effet de l'âge sur la longueur de l'aile (ou la coloration des pattes) par comparaison de mesures effectuées aux âges 1A – 2A – +2A sur les mêmes individus**



# Test de moyenne pour éch. pairés

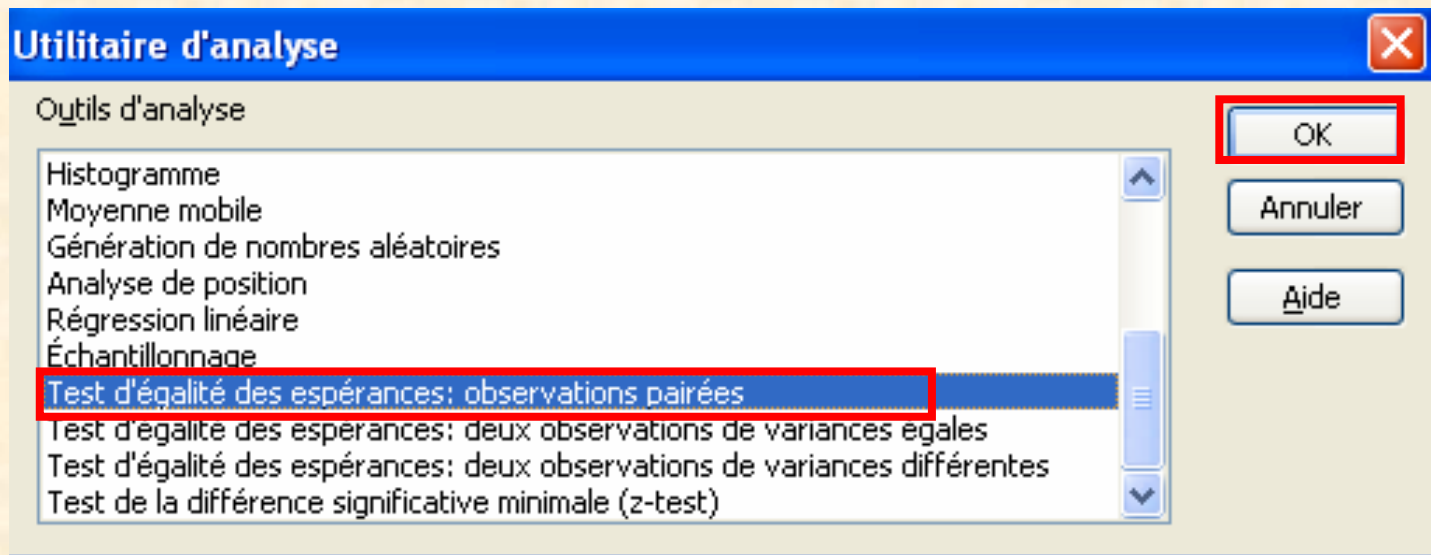
Etape 2: préparer les données au bon format pour l'Utilitaire d'analyse  
Une ligne correspond à 1 individu (l'unité de répétition)

	A	B	C
1		1A	+1A
2	Ind 1	73.0	74.7
3	Ind 2	71.0	71.0
4	Ind 3	72.5	73.0
5	Ind 4	70.5	71.0
6	Ind 5	70.5	71.2
7	Ind 6	71.0	71.3
8	Ind 7	72.5	74.2
9	Ind 8	71.5	72.3
10	Ind 9	72.5	74.1
11	Ind 10	68.5	68.9
12	Ind 11	74.0	75.6
13	Ind 12	70.0	70.2
14	Ind 13	70.5	71.2
15	Ind 14	71.0	71.2
16	Ind 15	72.0	73.2
17	Ind 16	73.0	73.2
18	Ind 17	68.5	70.0
19	Ind 18	75.0	76.3
20	Ind 19	72.5	73.7
21	Ind 20	72.0	73.5
22	Ind 21	69.5	71.1
23	Ind 22	72.5	73.1
24	Ind 23	74.0	75.4
25	Ind 24	70.0	71.4
26	Ind 25	71.0	71.7
27	Ind 26	70.5	71.0
28	Ind 27	70.0	71.1
29	Ind 28	70.5	71.2

Etape 3: Ouvrir l'Utilitaire d'analyse

et choisir

"Test d'égalité des espérances: observations pairées"



# Test de moyenne pour éch. pairés

## Etape 4: définition des données ET de l'analyse

	A	B	C
1		1A	+1A
2	Ind 1	73.0	74.7
3	Ind 2	71.0	71.0
4	Ind 3	72.5	73.0
5	Ind 4	70.5	71.0
6	Ind 5	70.5	71.2
7	Ind 6	71.0	71.3
8	Ind 7	72.5	74.2
9	Ind 8	71.5	72.3
10	Ind 9	72.5	74.1
11	Ind 10	68.5	68.9
12	Ind 11	74.0	75.6
13	Ind 12	70.0	70.2
14	Ind 13	70.5	71.2
15	Ind 14	71.0	71.2
16	Ind 15	72.0	73.2
17	Ind 16	73.0	73.2
18	Ind 17	68.5	70.0
19	Ind 18	75.0	76.3
20	Ind 19	72.5	73.7
21	Ind 20	72.0	73.5
22	Ind 21	69.5	71.1
23	Ind 22	72.5	73.1
24	Ind 23	74.0	75.4
25	Ind 24	70.0	71.4
26	Ind 25	71.0	71.7
27	Ind 26	70.5	71.0
28	Ind 27	70.0	71.1
29	Ind 28	70.5	71.2

**Test d'égalité des espérances: observations pairées**

**Paramètres d'entrée**

Plage pour la variable 1:

Plage pour la variable 2:

Différence entre les moyennes (hypothèse):

☒ Intitulé présent

Seuil de signification:

**Options de sortie**

☐ Plage de sortie:

☒ Insérer une nouvelle feuille:

☐ Créer un nouveau classeur

OK Annuler Aide

Définition de l' $H_0$  qui correspond à la question

0  $\leftrightarrow$  pas de différence  
2  $\leftrightarrow$  différence  
normalement attendue  
(d'après autres études)

# Test de moyenne pour éch. pairés

## Etape 5: interprétation des résultats

	1A	+1A
Moyenne	71.78125	72.7354231
Variance	3.33770161	4.2252481
Observations	32	32
Coefficient de corrélation de Pearson	0.96957972	
Différence hypothétique des moyennes	0	
Degré de liberté	31	
Statistique t	-10.1871474	
P(T<=t) unilatéral	1.0272E-11	
Valeur critique de t (unilatéral)	1.69551874	
P(T<=t) bilatéral	2.0544E-11	
Valeur critique de t (bilatéral)	2.03951344	