INTRODUCTION AUX STATISTIQUES NON PARAMETRIQUES

Sandrine Mignon-Grasteau

2014

QUAND FAUT IL UTILISER DES TESTS NON PARAMETRIQUES ?

- Non normalité des données => impossible d'utiliser les tests paramétriques

- On ne peut pas supposer que l'échantillon est tiré d'une distribution appartenant à une famille donnée

- Les variances sont hétérogènes

LES TESTS SONT QUAND MEME SENSIBLES AUX DONNEES ABERRANTES
MIEUX VAUT SE MEFIER EN CAS DE VARIANCES HETEROGENES

QUE FAIRE DANS CE CAS ?

- A : Rien, on fait du paramétrique quand même, et on espère que personne ne s'en apercevra
- B : on transforme les données pour les rendre normales, en espérant être capable d'interpréter les résultats obtenus sur l'échelle originale
- C : on fait du non paramétrique

QUE PEUT ON FAIRE ?

1. Comparaison d'un échantillon à une valeur
   A. Test du signe
   B. Test de Wilcoxon (signes + rangs)
   C. Test de Cox et Stuart (tendance)
   D. Test d'ajustement à une loi normale
2. Comparaison de 2 échantillons
   A. Test de la médiane
   B. Test de Wilcoxon-Mann-Whitney
   C. Test du signe ordonné de Wilcoxon (échantillons appariés)
   D. Test d'égalité des variances de Siegel-Tukey
3. Comparaisons de 3 échantillons ou plus
   A. Test de Kruskall-Wallis
   B. Test de Friedman (échantillons appariés)
   C. Homogénéité des variances
4. Corrélation entre données bivariates
   A. Corrélation de rang de Spearman
   B. Corrélations de rang de Kendall

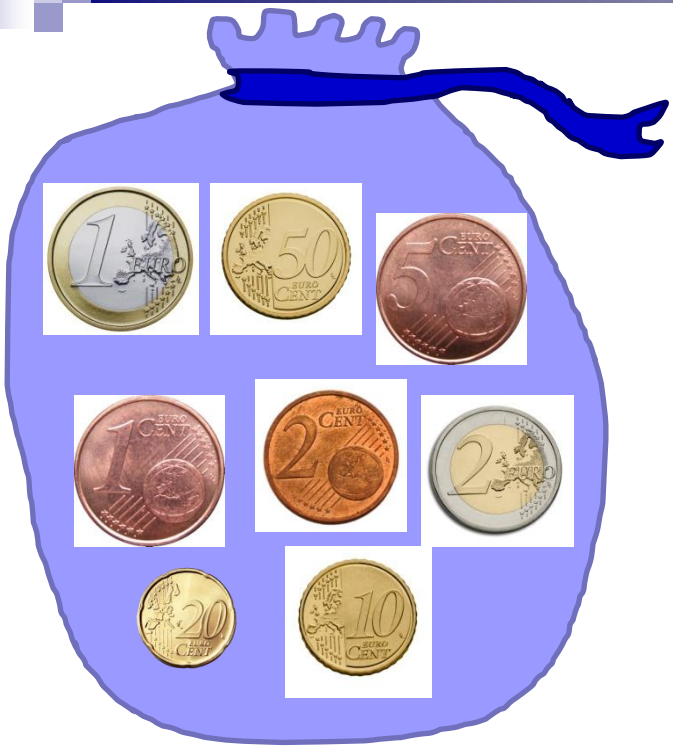
RAPPELS

- Factorielle :

$$n\,! = 1 \times 2 \times \ldots \times n$$

- Combinatoire : nombre de façons d'arranger p éléments dans un ensemble de n éléments

$$C_n^p = \frac{n!}{(n-p)! \times p!} = \frac{1 \times 2 \times \cdots \times n}{[1 \times 2 \times \cdots \times (n-p)] \times [1 \times 2 \times \cdots \times p]}$$



Combien de façons possibles ai-je de faire des combinaisons de 3 pièces parmi ces 8 pièces ?

$$C_8^2 = \frac{8!}{6!\quad 2!} = \frac{\cancel{1\times2\times3\times4\times5\times6}\times7\times\cancel{8}^{\,4}}{(\cancel{1\times2\times3\times4\times5\times6})\times(1\times\cancel{2})}$$

Je tire 9 pièces,
j'en obtiens 5 de 1 euro

Probabilité () = 1/8

Probabilité () = 7/8

1/8 × 1/8 × 1/8 × 1/8 × 1/8 × 7/8 × 7/8 × 7/8 × 7/8
=0.000018 … pour une combinaison

Et il y a C_9^5 combinaisons possibles, soit 126

La probabilité de tirer 5 pièces de 1 euro parmi 9 est donc :
Probabilité d'une combinaison ×nombre de combinaisons=
0.000018 ×126=0.00225
$= C_9^5 \times (1/8)^5 \times (7/8)^{9-5}$

La probabilité de tirer p pièces de 1 euro parmi n est donc :
$= C_n^p \times (p(1\text{ euro}))^p \times (p(\text{pas }1\text{ euro})^{n-p}$

COMPARAISON D'UN ECHANTILLON A UNE VALEUR

- **Test du signe**
- **Test de Wilcoxon (signes + rangs)**
- **Test de Cox et Stuart (tendance)**
- **Test d'ajustement à une loi normale**

1.A. TEST DU SIGNE

- Soit le poids de 12 cochons

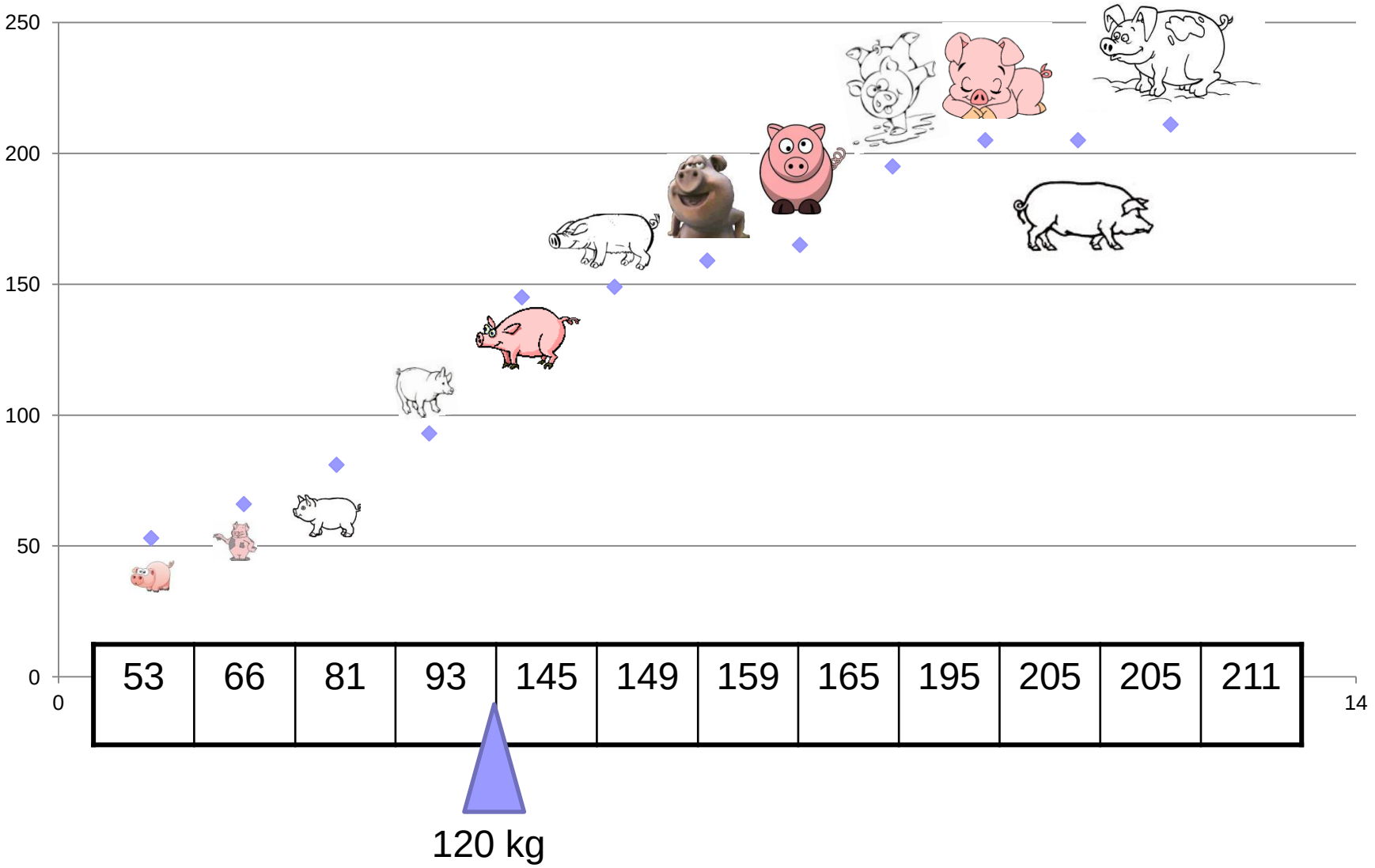
53	66	81	93	145	149	159	165	195	205	205	211

- On veut vérifier si la médiane est de 120 kg

- On peut se poser deux questions différentes :
  - H0 : $\Theta = 120$ contre H1a : $\Theta \neq 120$
  - H0 : $\Theta = 120$ contre H1b: $\Theta < 120$

-

Etape 1 : on regarde les données…



1.A. TEST DU SIGNE

- Soit le poids de 12 cochons

53	66	81	93	145	149	159	165	195	205	205	211

- On veut vérifier si la médiane est de 120 kg

- On peut se po
  - H0 : Θ = 1
  - H0 : Θ = 1

C'est à vous ...

- Sous H0, chaque cochon a la même probabilité d'être au-dessus ou au-dessous de 120 kg (6 cochons au dessus, 6 au dessous de la médiane)
  - ↳ la distribution des signes « + » suit donc une loi binomiale de paramètres n=12 et p=0.5

$$\mathbf{P(\text{nombre de signes "+"} = k) = C_n^k (p)^k (1-p)^{n-k}}$$

1.A. TEST DU SIGNE

- On compte le nombre d'observations au dessous de la médiane (-) ou au-dessus de la médiane (+) et on calcule la probabilité associée

53	66	81	93	145	149	159	165	195	205	205	211
-	-	-	-	+	+	+	+	+	+	+	+

Sous H0, la probabilité d'observer au maximum 4 signes « - » est égale à :

P(0 signe « - ») + P(1 signe « - ») + P(2 signes « - »)
+ P(3 signes « - ») + P(4 signes « - »)

1.A. TEST DU SIGNE

$$P(0 \text{ signe } "-") = C_{12}^{0}(0.5)^{0}(0.5)^{12} = \frac{12!}{0!\,12!}(0.5)^{12} = 0.00024$$

$$P(1 \text{ signe } "-") = C_{12}^{1}(0.5)^{1}(0.5)^{11} = \frac{12!}{1!\,11!}(0.5)^{12} = 0.0029$$

$$P(2 \text{ signes } "-") = C_{12}^{2}(0.5)^{2}(0.5)^{10} = \frac{12!}{2!\,10!}(0.5)^{12} = 0.0161$$

$$P(3 \text{ signes } "-") = C_{12}^{3}(0.5)^{3}(0.5)^{9} = \frac{12!}{3!\,9!}(0.5)^{12} = 0.0537$$

$$P(4 \text{ signes } "-") = C_{12}^{4}(0.5)^{4}(0.5)^{8} = \frac{12!}{4!\,8!}(0.5)^{12} = 0.1208$$

$P(H_0)$ = (0.0024+0.0029+0.0161+0.0537+0.1208) pour H1b
= 2*(0.0024+0.0029+0.0161+0.0537+0.1208) 0.388 pour H1a
> 0.05

↳ On ne rejette pas l'hypothèse H_0 selon laquelle la médiane des poids est égale à 120 kg

Pour que H0 soit rejetée au profit de H1, il faudrait au plus 2 cochons < la médiane

1.A. TEST DU SIGNE

- Approximation pour de grands échantillons : $Si \begin{cases} np \geq 5 \\ nq \geq 5 \end{cases}$

- On peut utiliser l'approximation suivante de la distribution de la variable : $Z = \frac{X - np}{\sqrt{npq}} \sim \aleph(0,1)$

C'est à vous ...

- **Ne pas utiliser ce test si n<5**

  - si n=5, le test peut être significatif si k=0 en unilatéral
  - si n=6-7, le test peut être significatif si k=0 en bilatéral
  - si n=8, le test peut être significatif si k=0, en bilatéral, si k=1 en unilatéral
  - si n=9-10, le test peut être significatif si k=0-1 en bilatéral …

Avec 5 individus : même la probabilité d'avoir 0 signe négatif est > 0.05 en bilatéral

$$P(0 \text{ signe } "-") = C_5^0 (0.5)^0 (0.5)^5 = \frac{5!}{0!\,5!}(0.5)^5 = 0.031$$

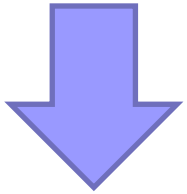
Avec 6 individus : seule la probabilité d'avoir 0 signe négatif est < 0.05 en bilatéral

$$P(0 \text{ signe } "-") = C_6^0 (0.5)^0 (0.5)^6 = \frac{6!}{0!\,6!}(0.5)^6 = 0.016$$

$$P(1 \text{ signe } "-") = C_6^1 (0.5)^0 (0.5)^6 = \frac{6!}{1!\,5!}(0.5)^6 = 0.078$$

Je veux toujours comparer un échantillon à une valeur théorique…

- … mais avec un peu plus d'élégance



Test de WILCOXON

1.B. TEST DE WILCOXON

- Combine l'information sur les signes et sur les rangs

53	66	81	93	145	149	160	165	195	205	207	211

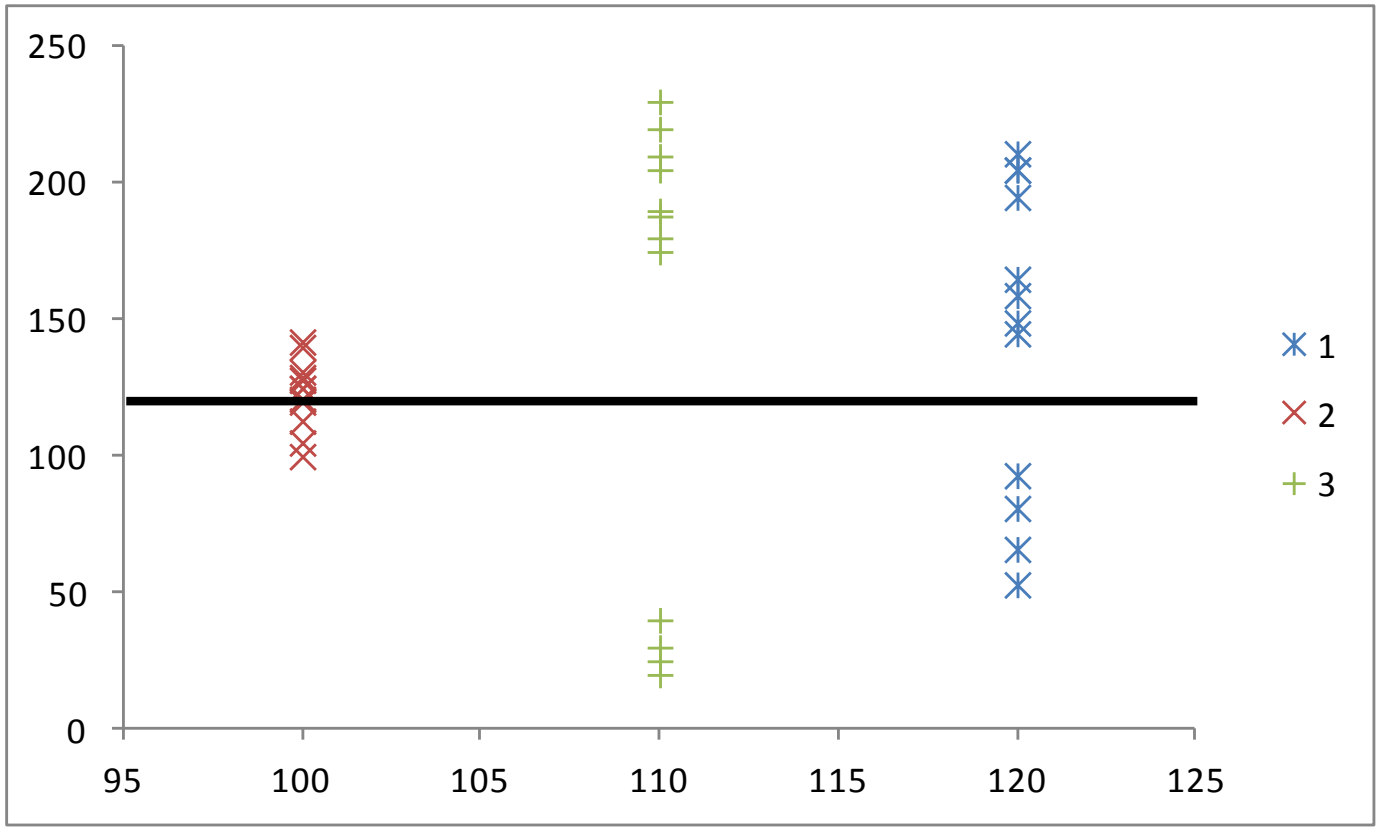
- On calcule la différence avec la médiane théorique (Θ=120)

-67	-54	-39	-27	25	29	40	45	75	85	87	91

- On les range par ordre croissant, en conservant les signes

25	27	29	39	40	45	54	67	75	85	87	91
+	-	+	-	+	+	-	-	+	+	+	+
1	2	3	4	5	6	7	8	9	10	11	12

Pourquoi cette combinaison rangs-signes ?



Avec le test des signes, ces 3 situations sont considérées comme identiques …

1.B. TEST DE WILCOXON

- On les range par ordre croissant, en conservant les signes

25	27	29	39	39	45	54	67	75	85	85	91
+	-	+	-	+	+	-	-	+	+	+	+
1	2	3	4	5	6	7	8	9	10	11	12

- On calcule la somme des rangs positifs (S_p) et la somme des rangs négatifs (S_n) :
  - $S_p = 1+3+5+6+9+10+11+12 = 57$
  - $S_n = 2+4+7+8 = 21$

- On consulte ensuite la table du test pour n=12, pour savoir si cet écart entre somme des rangs est significatif

1.B. TEST DE WILCOXON

n	$\alpha \leq 5\%$	$\alpha \leq 1\%$
6	0	-
7	2	-
8	4	0
9	6	2
10	8	3
11	1	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	20
17	35	23
18	40	28
19	46	32
20	52	38

- Dans cette table, pour n=12, le test n'est significatif que si la plus petite des deux sommes (S_n et S_p) est inférieure à 14

Ici, $\min(S_n,S_p)=21$

↳ H_0 n'est pas rejetée

1.B. TEST DE WILCOXON

- S'il y a des ex-aequo : on remplace le rang par un rang moyen
  - Ex : -5, 3, 3, 6, 8 ⇨ 1, 2.5, 2.5, 4, 5

- S'il y a des valeurs égales à la médiane
  - On attribue provisoirement le rang 1 aux valeurs nulles
    - Ex : 0, 3, -7, 11, -13, 16, -18 ⇨1, 2, 3, 4, 5, 6, 7
  - On multiplie les rangs par -1 pour les valeurs négatives, +1 pour les valeurs positives, 0 pour les valeurs nulles
    - Ex : 1, 2, 3, 4, 5, 6, 7 ⇨0, 2, -3, 4, -5, 6, -7
      - Sn = 15
      - Sp = 12

1.B. TEST DE WILCOXON

- Si n>20, on utilise une approximation normale

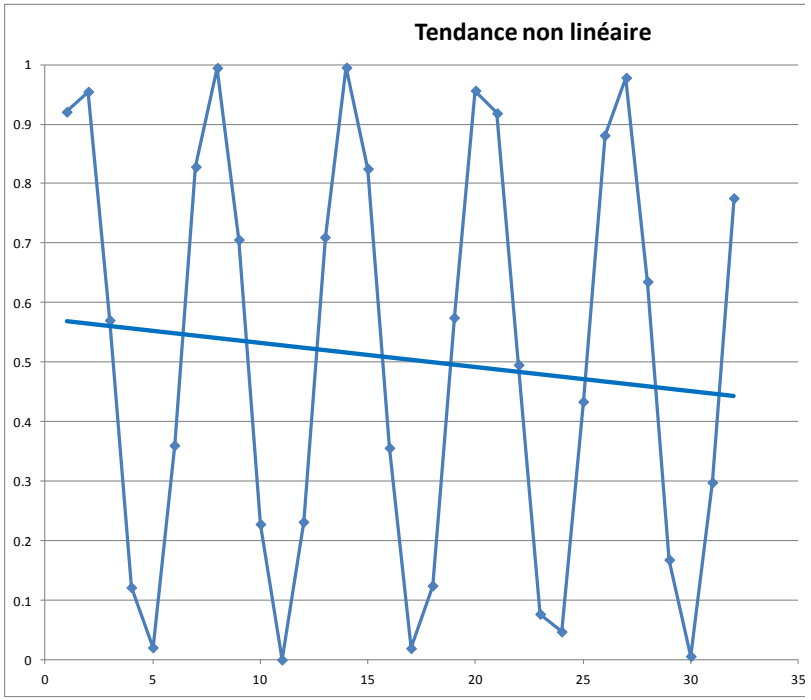
$$Z = \frac{S - \frac{1}{4}\left[n(n+1) - d_0(d_0+1)\right] - \frac{1}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{d_0(d_0+1)(2d_0+1)}{24} - \frac{\sum_i (d_i^3 - d_i)}{48}}} \sim \aleph(0,1)$$

- Avec d_0 = nombre de valeurs égales à la moyenne
- d_i = nombre de valeurs de rang i (s'il n'y a pas d'ex-aequo, tous les d_i valent 1)
- Sans ex-aequo ni valeurs égales à la médiane, Z se réduit à :

$$Z = \frac{S - \frac{1}{4}\left[n(n+1)\right]}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim \aleph(0,1)$$

- Je veux savoir si dans une série (chronologique, dose-dépendant, …), j'ai une tendance vers une évolution positive ou négative linéaire

Y= -0.0025 x+ 0.5711
R^2=0.007



Y= -0.004 x + 0.5734
R^2=0.0112

1.C. TEST DE COX-STUART

- C'est une adaptation du test des signes pour chercher une tendance monotone

- Soit n observations ordonnées dans le temps (de x_1 à x_{2m})

  - si n est pair, on calcule les différences :

    $$x_{m+1}-x_1 \ldots x_{2m}-x_m$$

  - Si n est impair, on supprime l'observation du milieu et on fait les mêmes calculs

1.C. TEST DE COX-STUART

H_0 : il n'y a pas de tendance (autant de différences positives que négatives, c.à.d. m)

H_1 : il y a une tendance (il y a plus de différences d'un signe que de l'autre)

Sous H_0, la probabilité d'avoir au plus X différences positives ou négatives est :

$$P = 2 \times \sum_{k=0,\ldots,X} C_m^k (0.5)^k (0.5)^{m-k}$$

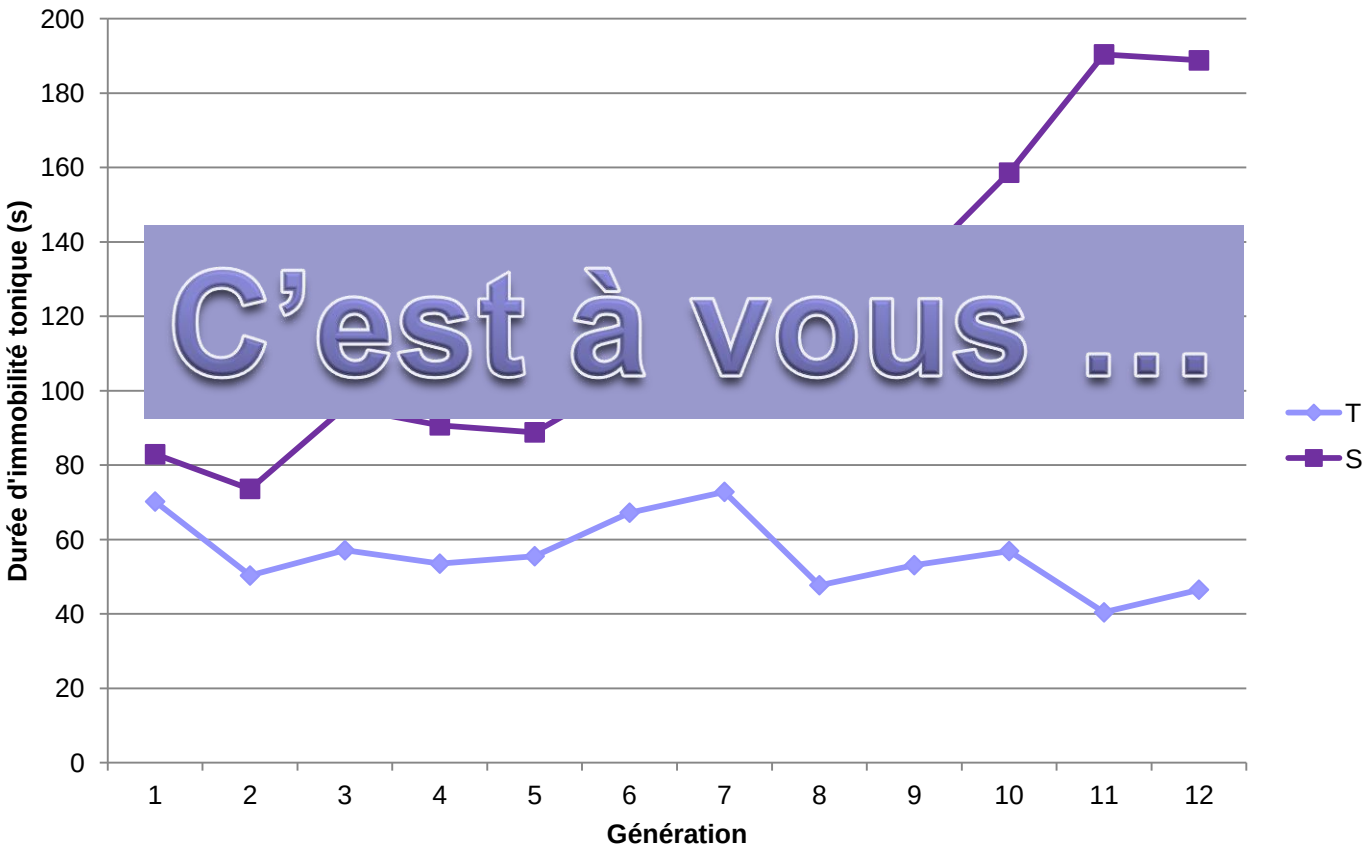
- **P> 0.05, H_0 est conservée, il n'y a pas de tendance**
- **P< 0.05, H_0 est rejetée au profit de H_1, il y a une tendance**

1.B. TEST DE COX-STUART

- Ex : Durée d'immobilité tonique chez la caille avec la génération de sélection (G) dans la lignée témoin (T) et la lignée sélectionnée (S)

G	1	2	3	4	5	6	7	8	9	10	11	12
T	70.2	50.3	57.1	53.5	55.5	67.2	72.8	47.7	53.1	56.9	40.4	46.5
S	82.9	73.6	95.5	90.7	88.8	103.5	129.2	115.2	132.8	158.6	190.4	188.8

Etape 1 : regarder les données …



1.C. TEST DE COX-STUART

G	1	2	3	4	5	6	7	8	9	10	11	12
T	70.2	50.3	57.1	53.5	55.5	67.2	72.8	47.7	53.1	56.9	40.4	46.5
S	82.9	73.6	95.5	90.7	88.8	103.5	129.2	115.2	132.8	158.6	190.4	188.8

	T	S
x7-x1	2.6	46.3
x8-x2	-2.6	41.5
x9-x3	-4.0	37.3
x10-x4	3.4	67.8
x11-x5	-15.1	101.6
x12-x6	-20.7	85.8

Nb. Différences	T	S
Négatives	4	0
Positives	2	6

1.C. TEST DE COX-STUART

Nb. Différences	T	S
Négatives	4	0
Positives	2	6

H0 **: il n'y a pas de tendance (autant de différences positives que négatives, càd 3)**
H1 **: il y a une tendance (il y a plus de différences d'un signe que de l'autre)**

Sous H0, la probabilité d'avoir au plus 2 différences positives ou négatives (T):

$$P = \sum_{k=0,1,2,4,5,6} C_6^k (0.5)^k (0.5)^{n-k} = 2\times(0.016+0.094+0.234) = 0.69$$

↳ P > 0.05, H0 n'est pas rejetée, il n'y a pas de tendance dans la lignée témoin

Sous H0, la probabilité d'avoir au plus 0 différence positives ou négatives (S):

$$P = \sum_{k=0,6} C_6^k (0.5)^k (0.5)^{n-k} = 0.03$$

↳ P < 0.05, H0 est rejetée au profit de H1, il y a une tendance dans la lignée sélectionnée

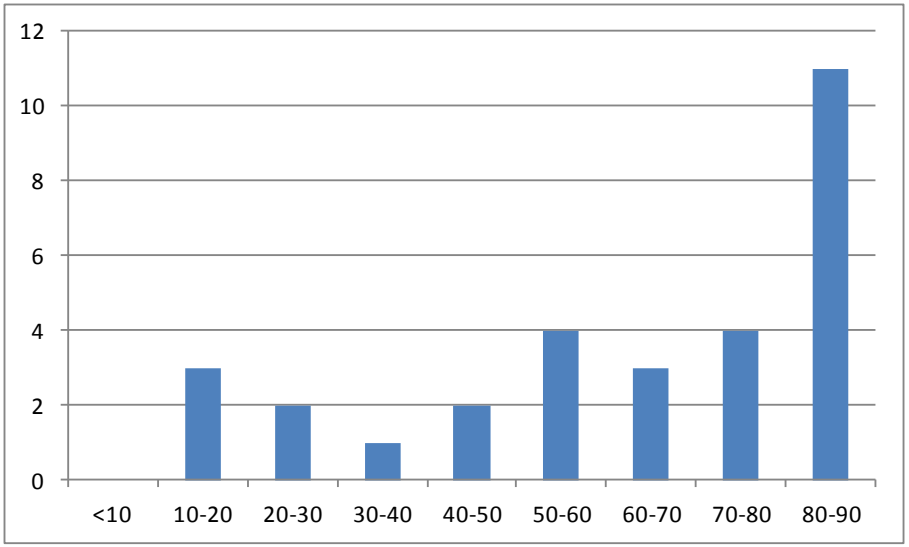
1.D. TEST DE LILLIEFORS

- Permet de tester si l'échantillon est issu d'une distribution normale
- Compare la répartition de la population et de la courbe de répartition empirique S_x :
  - S_x = (nb valeurs ≤ x dans l'échantillon)/n
- **Ex** : distribution de l'âge au décès

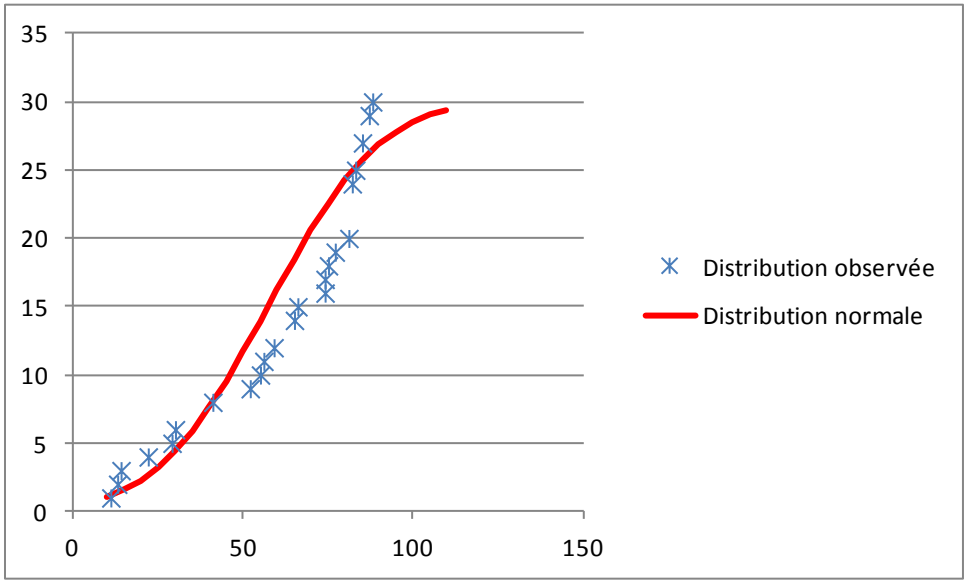
11	**13**	**14**	**22**	**29**	**30**	**41**	**41**	**52**	**55**	**56**	**59**	**65**	**65**	**66**
74	**74**	**75**	**77**	**81**	**82**	**82**	**82**	**82**	**83**	**85**	**85**	**87**	**87**	**88**

Etape 1 : on observe les données …

Nombre de morts observés par catégorie d'âge



Nombre de morts cumulés avec une distribution normale et avec la distribution observée



1.D. TEST DE LILLIEFORS

- On transforme les données pour les standardiser :
  - $z_i = (x_i - m)/s$ avec m=moyenne=61.43 et s=écart-type=25.04
- On compare la répartition empirique S à la répartition normale Φ sur les données standardisées
  - Ex : pour $x_1=11$, $z_1=-2.014$, et $\Phi(-2.014)=0.022$*

  * *$\Phi(-1.960)=0.025$*
- Pour chaque x_i, S est calculée comme 1/n
  - $S(z_1) = 1/30$
  - $S(z_2) = 1/30 + 1/30$
  - $S(z_7)= 1/30 + 1/30 + 1/30 + 1/30 + 1/30 + 1/30 + 2/30$ (ex-aequo)

1.D. TEST DE LILLIEFORS

x_i	z_i	$\Phi(z_i)$	$S(z_i)$	$\Phi(z_i)$-$S(z_i$-1)
			0	
11	-2.01	0.022	0.033	0.022
13	-1.93	0.026	0.067	-0.007
14	-1.89	0.029	0.100	-0.038
22	-1.57	0.058	0.133	-0.042
29	-1.29	0.098	0.167	-0.035
30	-1.25	0.105	0.200	-0.062
41*	-0.82	0.207	0.267	-0.007
52	-0.38	0.353	0.300	0.086
55	-0.26	0.399	0.333	0.099
56	-0.22	0.414	0.367	0.081
59	-0.10	0.461	0.400	0.094

x_i	z_i	$\Phi(z_i)$	$S(z_i)$	$\Phi(z_i)$-$S(z_i$-1)
65*	0.14	0.556	0.467	0.156
66	0.18	0.572	0.500	0.105
74*	0.50	0.692	0.567	**0.192**
75	0.54	0.706	0.600	0.139
77	0.62	0.733	0.633	0.133
81	0.78	0.782	0.667	0.149
82*	0.82	0.794	0.800	0.127
83	0.86	0.805	0.833	-0.005
85*	0.94	0.827	0.900	-0.006
87*	1.02	0.846	0.967	-0.054
88	1.06	0.856	1.000	-0.111

Max($\Phi(z_i)$-$S(z_{i-1})$) = 0.192 pour x=74, à comparer avec la valeur dans la table du test de Lilliefors

1.D. TEST DE LILLIEFORS : TABLE

N	6	7	8	9	10	11	12	13	14
5%	0.319	0.300	0.285	0.271	0.258	0.249	0.242	0.234	0.227
1%	0.364	0.348	0.331	0.311	0.294	0.284	0.275	0.268	0.261

N	15	16	17	18	19	20	25	30	>30
5%	0.220	0.213	0.206	0.200	0.195	0.190	0.173	0.161	$0.886/\sqrt{n}$
1%	0.257	0.250	0.245	0.239	0.235	0.231	0.203	0.187	$1.031/\sqrt{n}$

$\text{Max}(\Phi(z_i)-S(z_{i-1})) = 0.192$ pour x=74 ; Valeur critique pour n=30 : 0.161

↳On rejette l'hypothèse H_0 d'une distribution normale de l'âge du décès

COMPARAISON DE DEUX ECHANTILLONS

- Test de la médiane
- Test de Wilcoxon-Mann-Whitney
- Test de Siegel-Tukey
- Test d’égalité des variances

Je veux comparer deux groupes entre eux …

- Test de la médiane
- Test de Wilcoxon-Mann-Whitney
- Test de Siegel-Tukey

2.A. TEST DE LA MEDIANE

- ou Test de Mood
- Médiane équivalent de Moyenne en paramétrique
- Généralisation du test du signe au cas de deux échantillons indépendants
- H_0 : les deux échantillons ont la même médiane
  - M : médiane des deux échantillons regroupés
  - Sous H_0, il y a autant de valeurs supérieures à M dans chacun des deux échantillons, sinon il y a plus de valeurs supérieures à M dans l'un des deux échantillons
  - La distribution du nombre de valeurs supérieures à M = binomiale, p=0.50

2.A. TEST DE LA MEDIANE

- a_i = nb. d'observations >M dans l'échantillon i, de taille n_i

	Echantillon 1	Echantillon 2	Total
>M	a1	a2	$a1+a_2$
<M	n_1-a_1	$n2-a_2$	$n_1+n_2-a_1-a_2$
Total	n_1	n_2	$n1+n_2$

$$\mathbf{T = \frac{(2a_1 - n_1)^2 (n_1 + n_2)}{n_1 n_2} \sim \chi^2 \text{ à 1 ddl } *}$$

** Si n_1 et n_2 ne sont pas trop faibles*

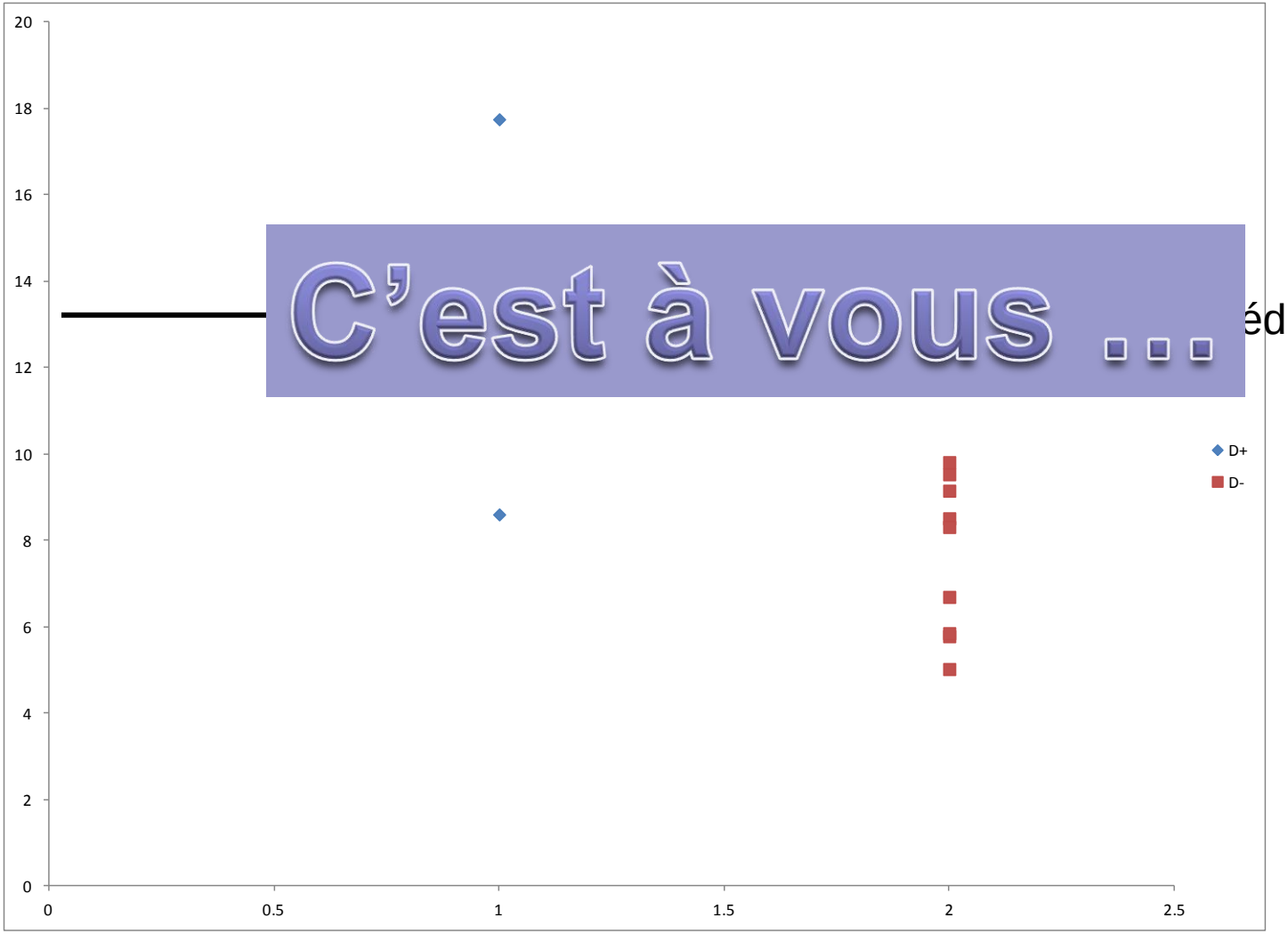
2.A. TEST DE LA MEDIANE

- **Lignées sélectionnées sur l'aptitude à digérer le blé (critère : Enérgie Métabolisable)**

D+	13.47	14.08	13.52	17.75	11.92	14.23	13.55	8.60	13.77	14.23	14.58	13.87
D-	5.78	6.69	9.53	12.82	5.02	14.04	5.85	8.51	11.53	9.81	8.31	9.15

- **M = 12.37**

Etape n°1 : on observe les données …



édiane = 12.37

2.A. TEST DE LA MEDIANE

- **Lignées sélectionnées sur l'aptitude à digérer le blé (critère : Enérgie Métabolisable)**

D+	13.47	14.08	13.52	17.75	11.92	14.23	13.55	8.60	13.77	14.23	14.58	13.87
D-	5.78	6.69	9.53	12.82	5.02	14.04	5.85	8.51	11.53	9.81	8.31	9.15

- **M = 12.37**

	D+	D-	Total
>M	10	2	12
<M	2	10	12
Total	12	12	24

2.A. TEST DE LA MEDIANE

	D+	D-	Total
>M	10	2	12
<M	2	10	12
Total	12	12	24

$$T = \frac{(2 \times 10 - 12)^2 (12 + 12)}{12 \times 12} = 10.67$$

- V.C. χ^2 à 1 d.d.l. = 3.84 (5%); 6.63 (1%); 10.83 (0.01%)
  - On rejette l'hypothèse H_0 selon laquelle les deux échantillons ont la même médiane, au seuil de 1%

Je veux toujours comparer deux échantillons entre eux…

- … mais j'utilise plus d'information que « au-dessus » ou « au-dessous de la moyenne »
- Le test de la médiane est au test de Mann-Whitney ce que le test du signe est au test de Wilcoxon

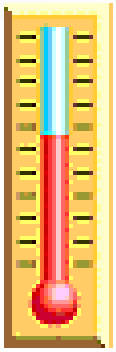
2.B. TEST DE WILCOXON-MANN-WHITNEY

- Wilcoxon (test de la somme des rangs) et Mann-Whitney ont élaborés leurs tests indépendamment
- Ecriture différente
- Principe identique :
  - on mélange deux échantillons et on compare les rangs. Si les deux échantillons proviennent de la même population, les rangs seront identiques.
  - Pour chaque échantillon i, on calcule la valeur :

    $$U_i = S_i - 0.5 n_i (n_i + 1)$$

    où S_i est la somme des rangs de l'échantillon i, de taille n_i
  - La plus petite des deux U_i est comparée à la valeur limite dans la table. H_0 est rejetée si $\min(U_1, U_2) < V.L.$

2.B. TEST DE WILCOXON-MANN-WHITNEY



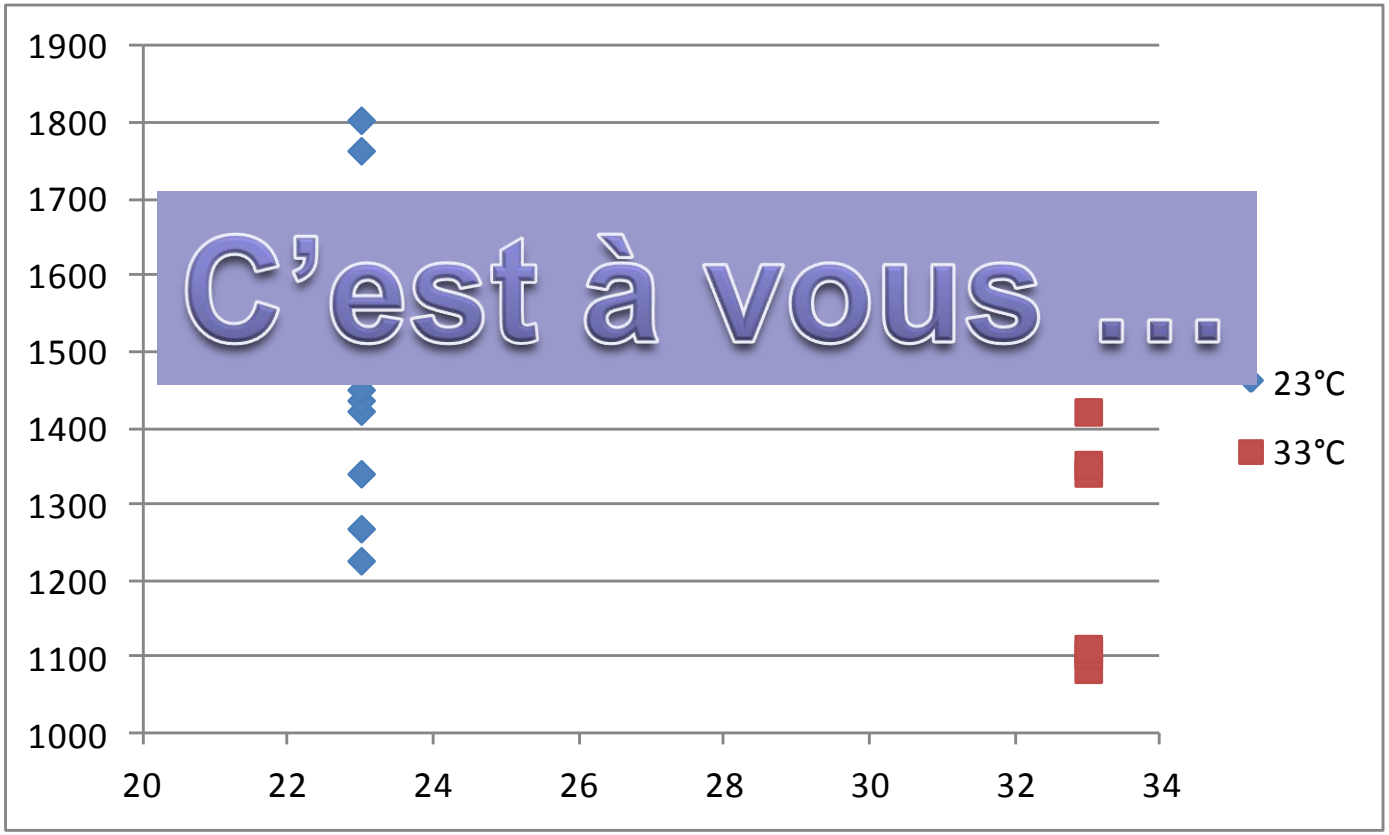
- Ex : poids des poulets élevés à 23°C ou 33°C

23°C	Rang	33°C	Rang
1227		1085	
1269		1104	
1341		1112	
1423		1342	
1437		1353	
1451		1422	
1507		1479	
1572		1534	
1584		1602	
1632			
1764			
1804			

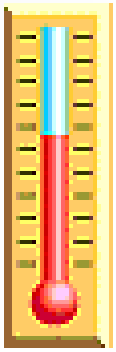
H_0 : les deux échantillons proviennent de la même population

H_1 : les deux échantillons proviennent de deux populations différant seulement par leur position (même variance)

Etape 1 : on regarde les données …



2.B. TEST DE WILCOXON-MANN-WHITNEY



- Ex : poids des poulets élevés à 23°C ou 33°C

23°C	Rang	33°C	Rang
1227	4	1085	1
1269	5	1104	2
1341	6	1112	3
1423	10	1342	7
1437	11	1353	8
1451	12	1422	9
1507	14	1479	13
1572	16	1534	15
1584	17	1602	18
1632	19		
1764	20		
1804	21		

H_0 : les deux échantillons proviennent de la même population

H_1 : les deux échantillons proviennent de deux populations différant seulement par leur position (même variance)

$U_{23°C} = 155 – 0.5 \times 12 \times 13 = 77$

$U_{33°C} = 76 – 0.5 \times 9 \times 10 = 31$

v.l. pour $n_1=12$ et $n_2=9$: 26 (5%)

Min(31,77) > 26

↳H_0 est conservée

2.B.TABLE DE MANN-WHITNEY (U)

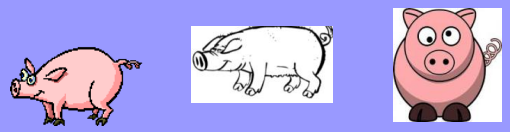
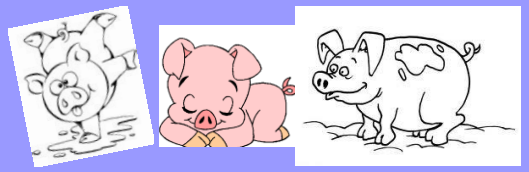
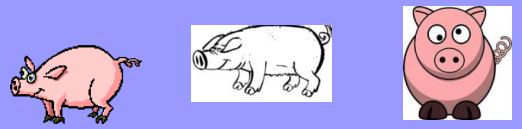
	n_1																			
n_2-n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	-	-	-	0	2	5	8	13	17	23	30	37	45	55	64	75	87	99	113	127
1	-	-	-	1	3	6	10	15	20	26	33	41	50	59	70	81	93	106	119	
2	-	-	0	2	5	8	12	17	23	29	37	45	54	64	75	86	99	112		
3	-	-	1	3	6	10	14	19	26	33	40	49	59	69	80	92	105			
4	-	-	1	4	7	11	16	22	28	36	44	53	63	74	85	98				
5	-	-	2	4	8	13	18	24	31	39	47	57	67	78	90					
6	-	0	2	5	9	14	20	26	34	42	51	61	72	83						
7	-	0	3	6	11	16	22	29	37	45	55	65	76							
8	-	0	3	7	12	17	24	31	39	48	58	69								
9	-	0	4	8	13	19	26	34	42	52	62									
10	-	1	4	9	14	21	28	36	45	55										
11	-	1	5	10	15	22	30	38	48											
12	-	1	5	11	17	24	32	41	50											
13	-	1	6	11	18	25	34	43												
14	-	1	6	12	19	27	36	45												
15	-	2	7	13	20	29	38													
16	-	2	7	14	22	30	40													
17	-	2	8	15	23	32														
18	-	2	8	16	24	33														
19	-	3	9	17	25															
20	-	3	9	17	27															
21	-	3	10	18																
22	-	3	10	19																
23	-	3	11																	
24	-	4	11																	
25	-	4																		

2.B. TEST DE MANN-WHITNEY (U)

- Approximation pour les grands échantillons (n, m > 20)
- U = min (U_n, U_m)
- $$z = \frac{U + \frac{1}{2} - \frac{1}{2}mn}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim \aleph(0,1)$$
- Dans l'exemple précédent z = -1.599 ⇨ p=0.11

Je veux toujours comparer deux groupes, mais …

- … mes échantillons ne sont plus indépendants les uns des autres
- Ils sont appariés

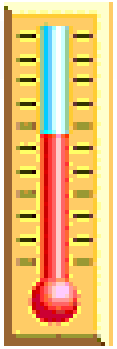
	Echantillons indépendants	Echantillons appariés
Condition 1	 *Naf-Naf* *Nif-Nif* *Nouf-Nouf*	 *Naf-Naf* *Nif-Nif* *Nouf-Nouf*
Condition 2	 *Riri* *Fifi* *Loulou*	 *Naf-Naf* *Nif-Nif* *Nouf-Nouf*

2.C. TEST DU SIGNE ORDONNE DE WILCOXON

- Echantillons appariés
- Les différences entre observations appariées constitue un ensemble unique qui ne peut être analysé avec les méthodes précédentes
- Soit x_i, y_i les deux mesures de l'individu i
- On calcule $d_i = x_i - y_i$
- H_0 : il n'y a pas de différence entre les x_i et les y_i, donc les d_i ont la même probabilité d'être positives ou négatives

↳ le test du signe est justifié

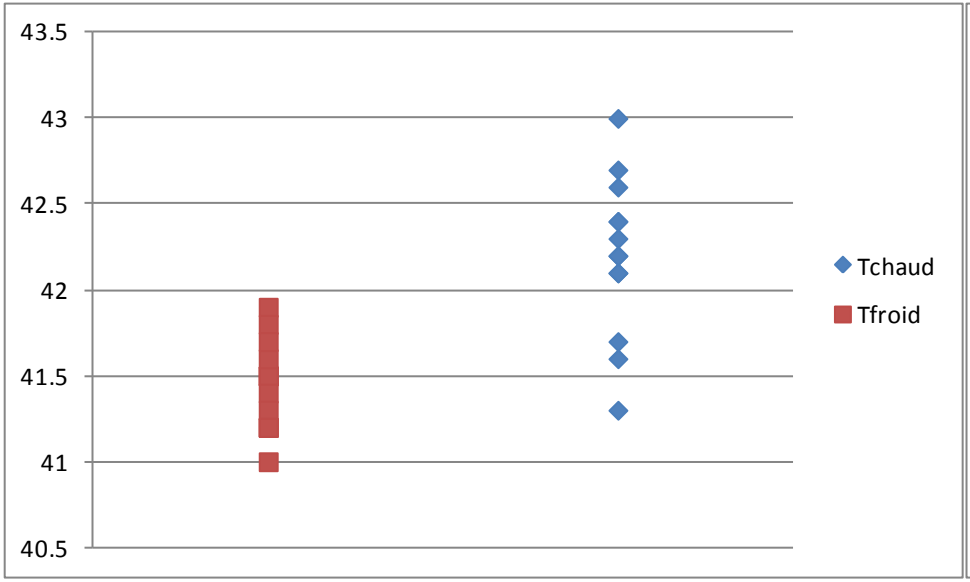
2.C. TEST DU SIGNE ORDONNE DE WILCOXON

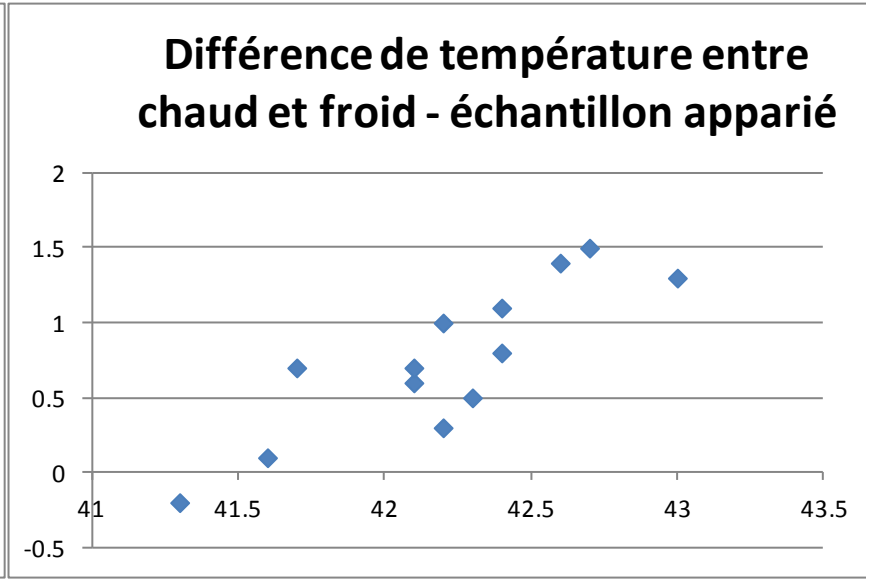


- Température des poulets mesurée au chaud ou au froid

Poulet	Tchaud	Tfroid	Tchaud-Tfroid	|Tchaud-Tfroid|	Rang	Signe
1	41.6	41.5	0.1	0.1	1	+
2	41.3	41.5	-0.2	0.2	2	-
3	42.2	41.9	0.3	0.3	3	+
4	42.3	41.8	0.5	0.5	4	+
5	42.1	41.5	0.6	0.6	5	+
6	42.1	41.4	0.7	0.7	6.5	+
7	41.7	41.0	0.7	0.7	6.5	+
8	42.4	41.6	0.8	0.8	8	+
9	42.2	41.2	1.0	1.0	9	+
10	42.4	41.3	1.1	1.1	10	+
11	43.0	41.7	1.3	1.3	11	+
12	42.6	41.2	1.4	1.4	12	+
13	42.7	41.2	1.5	1.5	13	+

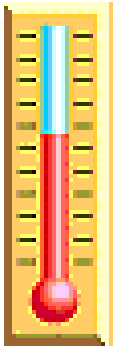
Etape 1 : on regarde les données …





Ce graphique-là, on ne peut le faire qu'avec des échantillons appariés

2.C. TEST DU SIGNE ORDONNE DE WILCOXON



- Température des poulets mesurée au chaud ou au froid

Rang	Signe
1	+
2	-
3	+
4	+
5	+
7	+
7	+
8	+
9	+
10	+
11	+
12	+
13	+

Test du signe :

$$\mathbf{P(0\ ou\ 13\ signes\ négatifs) = 2 \times C_{13}^{0} \times 0.5^{13} = 0.00024}$$

$$\mathbf{P(1\ ou\ 12\ signes\ négatifs) = 2 \times C_{13}^{1} \times 0.5^{13} = 0.00317}$$

$$\mathbf{\Rightarrow P = 0.003}$$

$$\mathbf{\Rightarrow H_0\ rejetée}$$

Test de Wilcoxon-Mann-Whitney :

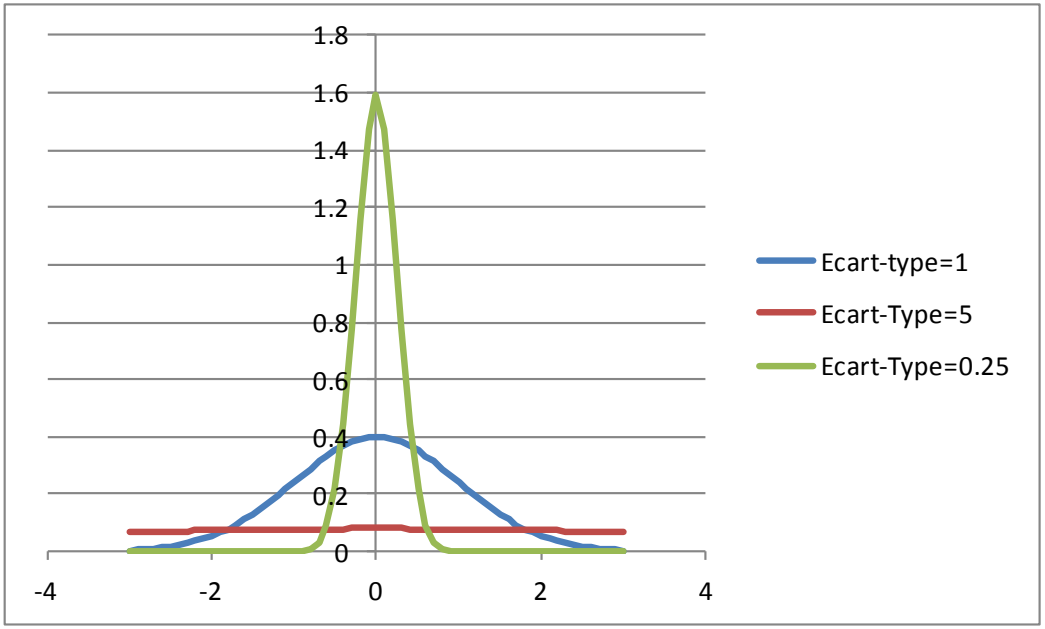
$S_n = 2 < 10$ (v.l. à 1% pour n=13)

2.C. TEST DU SIGNE ORDONNE DE WILCOXON : TABLE

n	$\alpha \leq 5\ \%$	$\alpha \leq 1\ \%$
6	0	-
7	2	-
8	4	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	20
17	35	23
18	40	28
19	46	32
20	52	38

Tester l'égalité de variances entres groupes

- Pourquoi tester l'égalité des variances entre groupes ?



Parce que ces 3 distributions, de même moyenne, sont différentes …

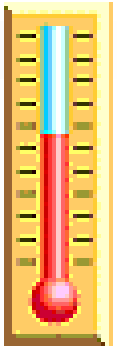
2.D. TEST D'EGALITE DES VARIANCES DE SIEGEL-TUKEY

- Équivalent du test paramétrique de Fisher-Snedecor

- Si les deux populations considérées ne diffèrent que par la variance, l'une des deux va avoir les positions les plus éloignées de la médiane, l'autre les positions les plus proches de la médiane

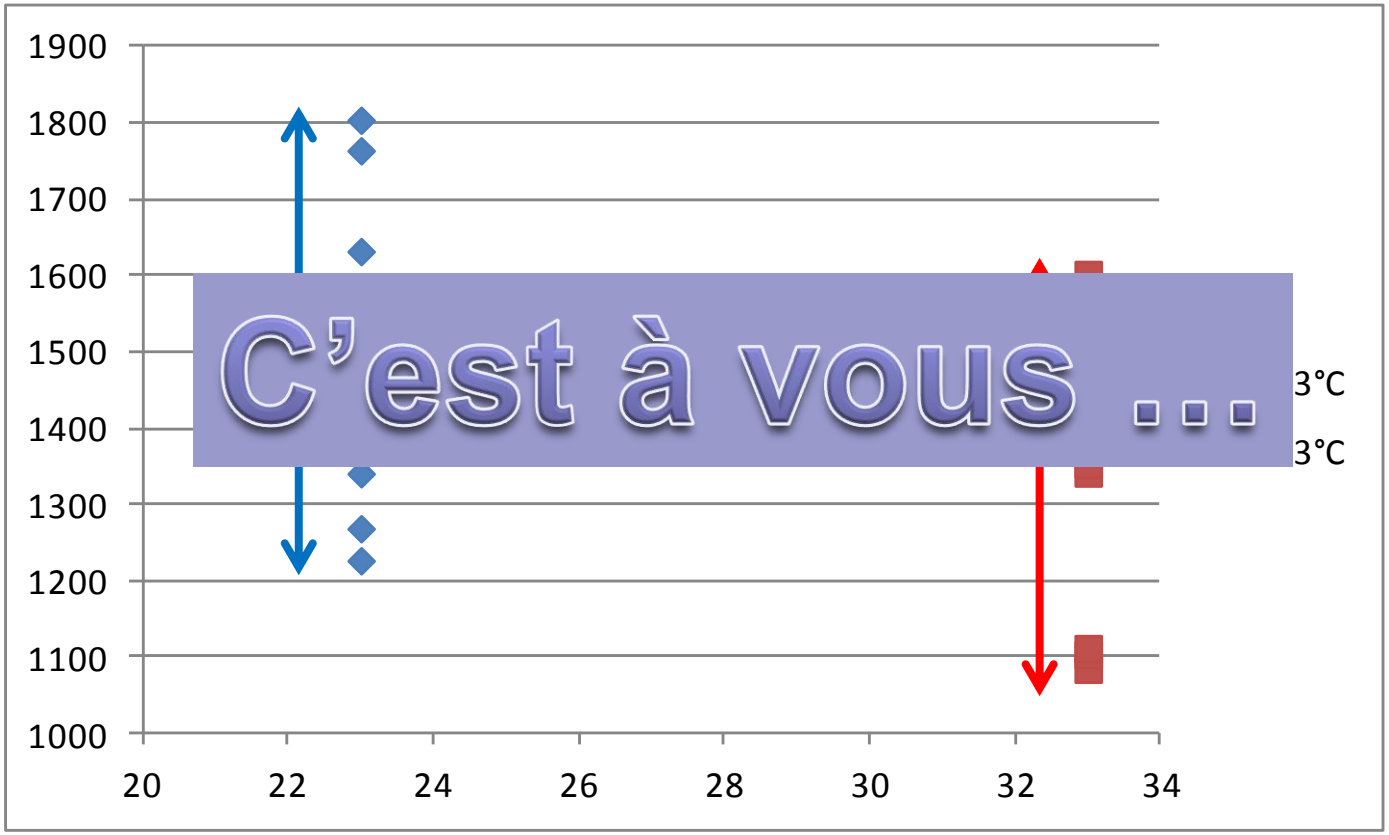
- On range l'échantillon regroupé par ordre croissant et on attribue la valeur 1 au plus petit, 2 au plus grand, 3 au 2ème plus petit, 4 au 2ème plus grand …
  - Ex : pour les valeurs 8, 12, 25, 38, 54 on attribue les rangs 1, 3, 5, 4, 2

2.D. TEST D'EGALITE DES VARIANCES DE SIEGEL-TUKEY

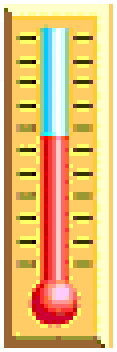


23°C	Rang	33°C	Rang
1227		1085	
1269		1104	
1341		1112	
1423		1342	
1437		1353	
1451		1422	
1507		1479	
1572		1534	
1584		1602	
1632			
1764			
1804			

Etape 1 : on regarde les données …



2.D. TEST D'EGALITE DES VARIANCES DE SIEGEL-TUKEY



23°C	Rang	33°C	Rang
1227	7	1085	1
1269	9	1104	3
1341	11	1112	5
1423	19	1342	13
1437	21	1353	15
1451	20	1422	17
1507	16	1479	18
1572	12	1534	14
1584	10	1602	8
1632	6		
1764	4		
1804	2		

- *Test de Siegel-Tukey* :
- $S_{23°C} = 137$, $U_{23°C} = 59$
- $S_{33°C} = 94$, $U_{33°C} = 49$
- v.l. pour n=9, m=12 : 45
  - H_0 conservée

- *Test de Fisher Snedecor :*
- $V_{23°C} = 32480.45$
- $V_{33°C} = 38140.25$
- $F = 1.174 < 3.31$

COMPARAISON DE TROIS ECHANTILLONS OU PLUS

- Test de Kruskall-Wallis
- Test de Friedman (échantillons appariés)
- Homogénéité des variances

Rapprochons nous de la vraie vie

- On n'a rarement que deux situations-groupes-traitements à comparer
- On a les mêmes questions qu'avant, mais avec plus de 2 groupes
  - Mes x groupes sont-ils équivalents
    - (groupes indépendants, Kruskall-Wallis)
    - (groupes appariés, Friedman)
  - Mes x groupes ont-ils la même variance ?

3.A. TEST DE KRUSKALL-WALLIS

- Généralisation du test de Wilcoxon-Mann-Whitney pour plus de deux échantillons
- On a k traitements sur lesquels on fait plusieurs observations **indépendantes**
- H_0 : les k échantillons ont la même distribution
- H_1 : au moins un des k échantillons n'a pas la même distribution
- On mélange les k échantillons et on remplace les observations par leur rang. On calcule ensuite la statistique de test H :

$$H = \left[\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{(R_j)^2}{n_j} \right] - 3(N+1)$$

  - R_j = somme des rangs de l'échantillon j
  - nj : effectif de l'échantillon j
  - N = effectif global

Si $n_j \leq 5$ ⇨Table
Si $n_j > 5$ ⇨ H ~ χ^2 à (k-1) d.d.l.

3.A. TEST DE KRUSKALL-WALLIS

- Ex : on compare les notes attribuées par 3 juges sur sept verrats (chaque juge note des verrats différents = indépendants)

Juge 1	Juge 2	Juge 3
65	77	79
67	69	83
80	66	78
72	85	92
75	89	94
57	74	70
76	60	90

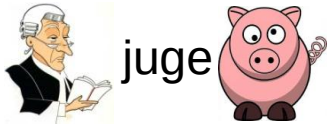
Etape 1 : on observe les données …


juge

et sa famille


juge

et sa famille


juge

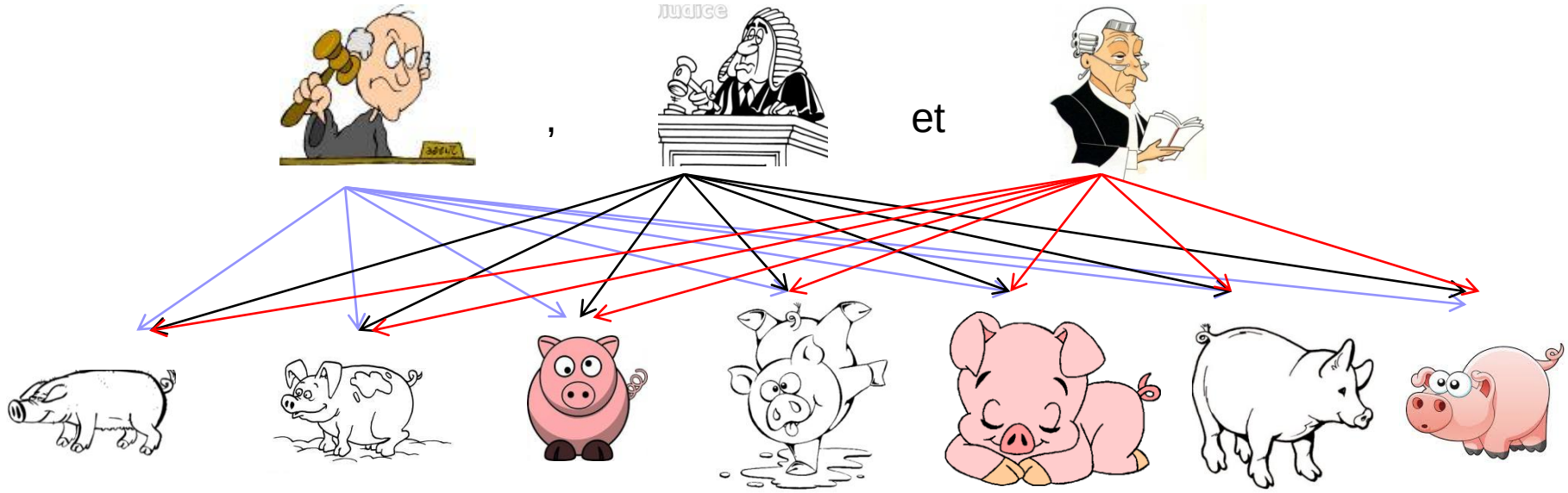
et sa famille

3.A. TEST DE KRUSKALL-WALLIS

Juge 1	Rang	Juge 2	Rang	Juge 3	Rang
65	3	77	12	79	14
67	5	69	6	83	16
80	15	66	4	78	13
72	8	85	17	92	20
75	10	89	18	94	21
57	1	74	9	70	7
76	11	60	2	90	19
$\boldsymbol{R_j}$	***53***		***68***		***110***

H = 6.479 à comparer avec un χ^2 à 2 d.d.l. (v.l. 5% = 5.991)

On prend les mêmes et on recommence …mais avec des échantillons appariés



jugent tous les 3 les mêmes cochons

3.B. TEST DE FRIEDMAN

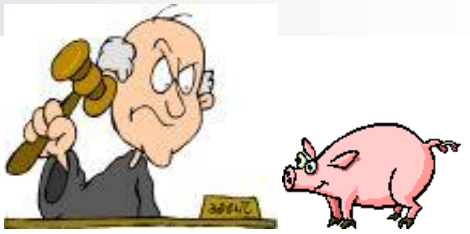
- Echantillons appariés (on fait subir plusieurs traitements à un même sujet)
- H_0 : les traitements ne diffèrent pas entre eux (en tenant compte de l'appariement)
- H_1 : au moins un des traitements diffère
- x_{ij} : observation du sujet i (i=1, n) avec le traitement j (j=1,k)
- On remplace l'observation x_{ij} par son rang pour le sujet i, et on calcule la somme des rangs pour chaque traitement j
- On calcule ensuite S :

$$S = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3n(k+1) \sim \chi^2_{k-1}$$

n : nombre de sujets ; k : nombre de traitements ;

R_j : somme des rangs pour le traitement j

3.B. TEST DE FRIEDMAN



- Ex : Si on reprend l'exemple précédent, les 3 juges notent les mêmes verrats

Verrat	Juge 1	Juge 2	Juge 3
1	65	77	79
2	67	69	83
3	80	66	78
4	72	85	92
5	75	89	94
6	57	74	70
7	76	60	90

Etape 1 …

3.B. TEST DE FRIEDMAN

- On remplace les valeurs par leurs rangs intra-sujet

Verrat	Juge 1	Juge 2	Juge 3
1	1	2	3
2	1	2	3
3	3	1	2
4	1	2	3
5	1	2	3
6	1	3	2
7	2	1	3
Total	10	13	19

$$S = \left[\frac{12}{21(3+1)}(10^2 + 13^2 + 19^2)\right] - 21(3+1) = 6.00 > 5.99 \text{ (v.l. de } \chi^2_2 \text{ à 5\%)}$$

↳ La différence entre juges est juste significative

3.C. HOMOGENEITE DES VARIANCES

- Généralisation du test des carrés des rangs
- On remplace les observations par les valeurs absolues des écarts à la moyenne dans chaque échantillon, on ordonne ces écarts absolus en combinant tous les échantillons, puis on calcule le carré de leurs rangs

3.C. HOMOGENEITE DES VARIANCES

Juge 1	Ecart à m	Juge 2	Ecart à m	Juge 3	Ecart à m
65	5.29	77	2.71	79	4.71
67	3.29	69	5.29	83	0.71
80	9.71	66	8.29	78	5.71
72	1.71	85	10.71	92	8.29
75	4.71	89	14.71	94	10.29
57	13.29	74	0.29	70	13.71
76	5.71	60	14.29	90	6.29
m=70.29		***m=74.29***		***m=83.71***	

3.C. HOMOGENEITE DES VARIANCES

Juge 1		**Juge 2**		**Juge 3**	
Ecart à m	**Rang (Rang²)**	**Ecart à m**	**Rang (Rang²)**	**Ecart à m**	**Rang (Rang²)**
5.29	8.5(72.25)	2.71	4(16)	4.71	6.5(42.25)
3.29	5(25)	5.29	8.5(72.25)	0.71	2(4)
9.71	15(225)	8.29	13.5(182.25)	5.71	10.5(110.25)
1.71	3(9)	10.71	17(289)	8.29	13.5(182.25)
4.71	6.5(42.25)	14.71	21(441)	10.29	16(256)
13.29	18(324)	0.29	1(1)	13.71	19(361)
5.71	10.5(110.25)	14.29	20(400)	6.29	12(144)
s_1	***807.75***	s_2	***1401.5***	s_3	***1099.85***

3.C. HOMOGENEITE DES VARIANCES

$$\sum \text{rangs}_{ij}^2 = 3309.1$$

$$C = \frac{(\sum \text{rangs}_{ij}^2)^2}{N} = \frac{3309.1^2}{21} = 521403.86$$

En absence d'ex – aequo :

$$S_t^2 - C = \frac{(N-1)N(N+1)(2N+1)(8N+11)}{180}$$

$$S_j^2 = \frac{807.75^2 + 1401.5^2 + 1099.85^2}{7} = 546618.91$$

$$H = (N-1)\frac{S_j^2 - C}{S_t^2 - C} = 1.28 \sim \chi^2 \text{ à } 2 \text{ d.d.l.}$$

↳ La différence de variance entre juges n'est pas significative

Dans tous les cas …

- **Que l'on compare les médianes ou les variances des n groupes, le test dit « il y en a au moins un qui est différent », mais pas « c'est celui-là qui est différent des autres »**
- **On doit ensuite refaire les tests 2 à 2 pour savoir qui est différent de qui**
- **Il faut alors penser à la correction de Bonferroni (on divise la probabilité a par le nombre de tests)**
  - Pour nos 3 groupes de cochons, on compare J1 et J2, J1 et J3, J2 et J3
  - On fait 3 tests
  - Pour déclarer une de ces différences significatives, elle doit être significative au seuil de 0.05/3=0.017 et non pas au seuil habituel de 0.05
  - On n'est pas obligés de faire tous les tests …

CORRELATIONS ENTRE DONNEES BIVARIATES

- Corrélation de rang de Spearman
- Corrélations de rang de Kendall

4.A. CORRELATION DE RANGS DE SPEARMAN

- Equivalent du coefficient de corrélation classique, mais calculé à partir des rangs et non des observations
- Utilise deux séries appariées X et Y, constituant n couples (X, Y)
- On calcule les rangs pour X, puis pour Y
- On calcule ensuite, pour chaque sujet i, la différence entre son rang pour X et son rang pour Y (d_i)
- On calcule enfin le coefficient de corrélation r_{XY} comme suit :

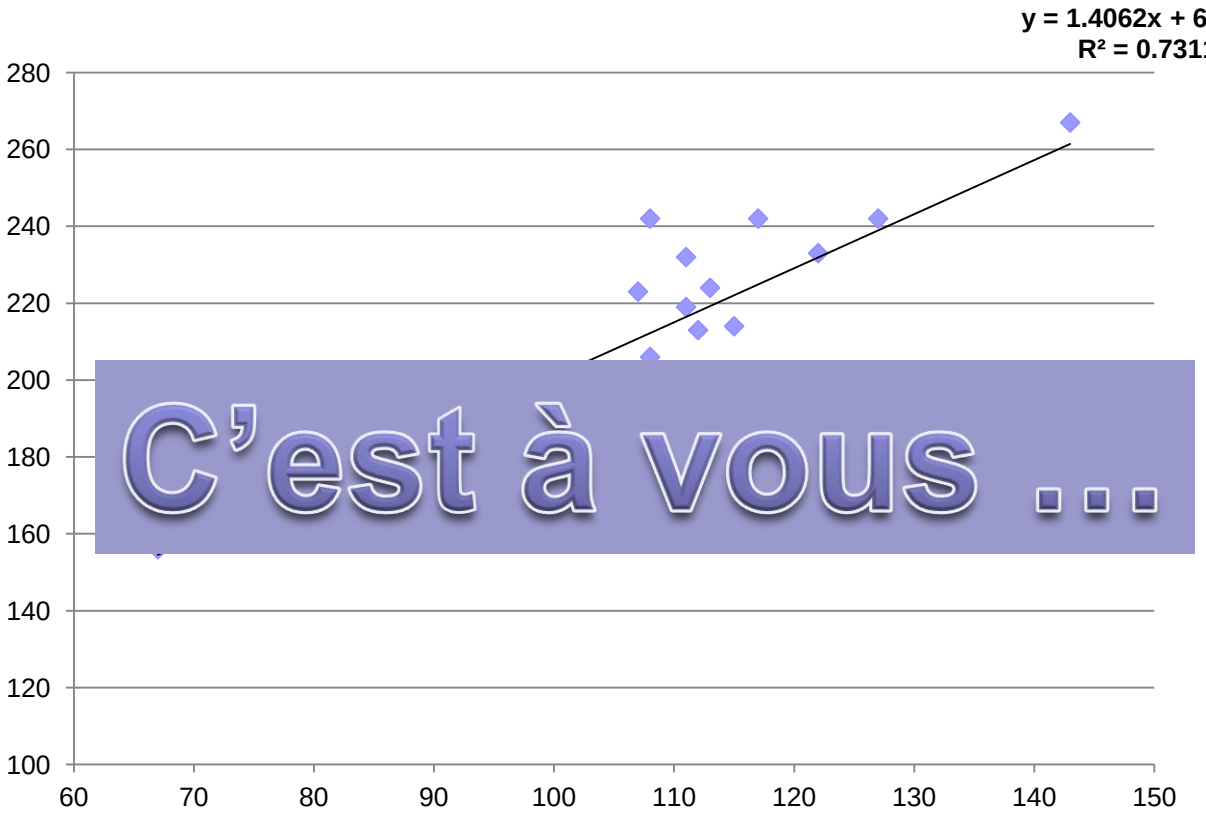
$$r_{XY} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

4.A. CORRELATION DE RANGS DE SPEARMAN

Sujet	Filet	Cuisse
1	111	232
2	143	267
3	113	224
4	122	233
5	94	200
6	111	219
7	82	178
8	111	192
9	107	189
10	108	206
11	108	242
12	115	214
13	127	242
14	112	213
15	117	242
16	80	178
17	107	223
18	105	185
19	111	196
20	67	156

- Les poulets qui ont les plus gros filets ont-ils les plus grosses cuisses ?

Etape 1 …



C’est à vous …

4.A. CORRELATION DE RANGS DE SPEARMAN

Sujet	Filet	Cuisse
1	111	232
2	143	267
3	113	224
4	122	233
5	94	200
6	111	219
7	82	178
8	111	192
9	107	189
10	108	206
11	108	242
12	115	214
13	127	242
14	112	213
15	117	242
16	80	178
17	107	223
18	105	185
19	111	196
20	67	156

- Les poulets qui ont les plus gros filets ont-ils les plus grosses cuisses ?

- ***Etape 2*** : calculer les rangs des animaux pour la variable « filet » et pour la variable « cuisse »

4.A. CORRELATION DE RANGS DE SPEARMAN

Sujet	Filet	Cuisse	Rang Filet	Rang Cuisse
1	111	232	11	15
2	143	267	20	20
3	113	224	15	14
4	122	233	18	16
5	94	200	4	8
6	111	219	12	12
7	82	178	3	3
8	111	192	13	6
9	107	189	6	5
10	108	206	8	9
11	108	242	9	17
12	115	214	16	11
13	127	242	19	19
14	112	213	14	10
15	117	242	17	18
16	80	178	2	2
17	107	223	7	13
18	105	185	5	4
19	111	196	10	7
20	67	156	1	1

- ***Etape 3*** : calculer les différences entre les rangs des animaux pour la variable « filet » et pour la variable « cuisse » et son carré

4.A. CORRELATION DE RANGS DE SPEARMAN

	Filet	Cuisse	Rang Filet	Rang Cuisse	d_i	d^2_i
1	67	156	1	1	0	0
2	80	178	2	2.5	-0.5	0.25
3	82	178	3	2.5	0.5	0.25
4	94	200	4	8	-4	16
5	105	185	5	4	1	1
6	*107*	*189*	6.5	5	1.5	2.25
7	*107*	*223*	6.5	13	-6.5	42.25
8	108	206	8.5	9	-0.5	0.25
9	108	242	8.5	18	-9.5	90.25
10	111	232	11.5	15	-3.5	12.25
11	111	219	11.5	12	-0.5	0.25
12	111	192	11.5	6	5.5	30.25
13	111	196	11.5	7	4.5	20.25
14	112	213	14	10	4	16
15	113	224	15	14	1	1
16	115	214	16	11	5	25
17	117	242	17	18	-1	1
18	122	233	18	16	2	4
19	127	242	19	18	1	1
20	143	267	20	20	0	0

- ***Etape 2*** : calculer les rangs des animaux pour les variables « filet » et « cuisse »
- ***Etape 3*** : calculer les différences de rang pour les deux variables
- $\Sigma d_i^2 = 263.5$
- ***Etape 4***: calculer la corrélation

$$r_{Filet,Cuisse} = 1 - \frac{6 \times 263.5}{20 \times (400 - 1)} = 0.80$$

4.A. CORRELATION DE RANGS DE SPEARMAN

- ***Etape 5*** : la corrélation est-elle significativement différente de 0 ?
- Si n<10 ⇨ Consulter la table

n	5	6	7	8	9	10
$\alpha \leq 5\%$	1.00	0.89	0.79	0.74	0.68	0.65
$\alpha \leq 1\%$	-	1.00	0.93	0.88	0.83	0.79

- Si n≥10 ⇨ la corrélation suit approximativement une loi normale :

$$t = \frac{r_{X,Y}}{\sqrt{1 - r_{X,Y}^2}} \times \sqrt{n-2} \sim t \text{ à } (n-2) \text{ d.d.l.}$$

C'est à vous ...

4.B. CORRELATION DE KENDALL

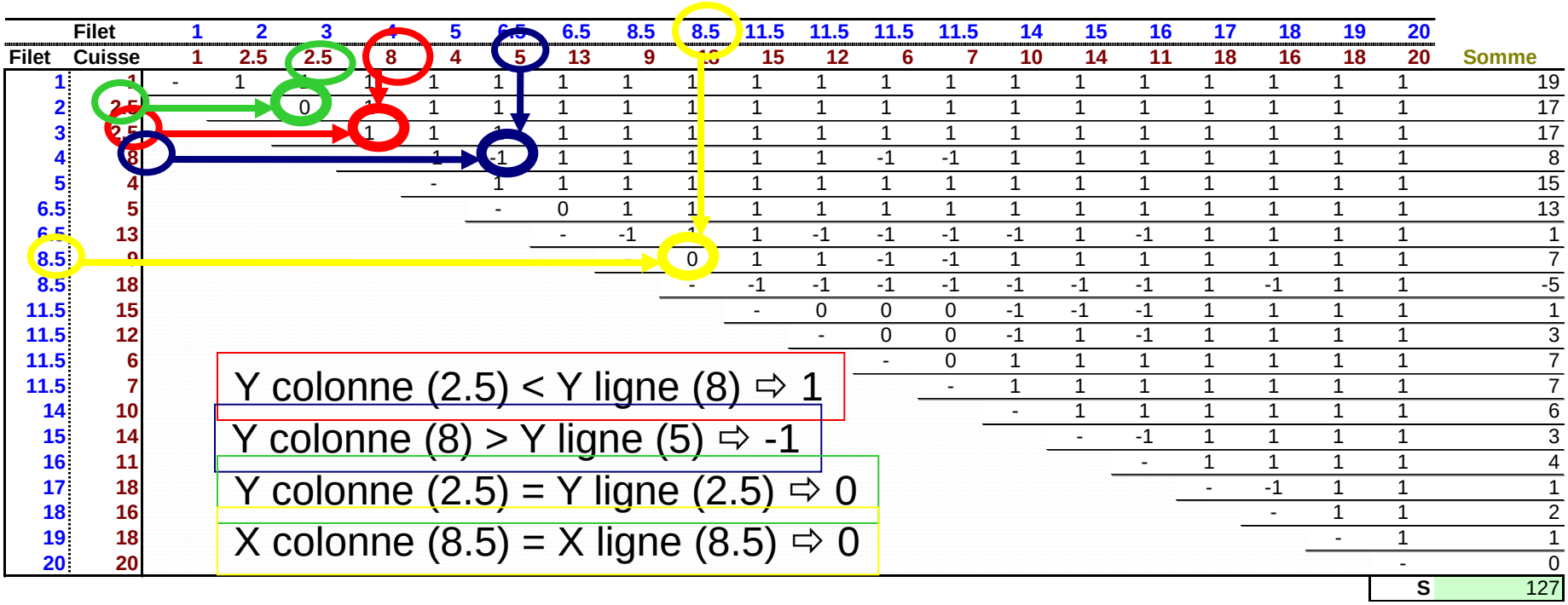
- *On ordonne sur le rang de la variable X*
- *On calcule les rangs pour la variable Y*
- *On construit un tableau avec les rangs de X et Y en ligne et en colonne*
- *Pour chaque observation, on compare les rangs de Y en ligne et en colonne*
  - *Rang Y en colonne < Rang Y en ligne => 1*
  - *Rang Y en colonne > Rang Y en ligne => -1*
  - *Rang Y en colonne = Rang Y en ligne => 0*
  - *Rang X en colonne = Rang X en ligne => 0*
  - *On fait le total des valeurs => S de la formule*
- *Le coefficient τ de Kendall vaut alors :*

$$\tau = \frac{2 \times S}{n(n-1)}$$

4.B. CORRELATION DE KENDALL

	Filet	*Cuisse*	*Rang Filet*	*Rang Cuisse*
1	**67**	**156**	**1**	**1**
2	**80**	**178**	**2**	**2.5**
3	**82**	**178**	**3**	**2.5**
4	**94**	**200**	**4**	**8**
5	**105**	**185**	**5**	**4**
6	***107***	***189***	**6.5**	**5**
7	***107***	***223***	**6.5**	**13**
8	**108**	**206**	**8.5**	**9**
9	**108**	**242**	**8.5**	**18**
10	**111**	**232**	**11.5**	**15**
11	**111**	**219**	**11.5**	**12**
12	**111**	**192**	**11.5**	**6**
13	**111**	**196**	**11.5**	**7**
14	**112**	**213**	**14**	**10**
15	**113**	**224**	**15**	**14**
16	**115**	**214**	**16**	**11**
17	**117**	**242**	**17**	**18**
18	**122**	**233**	**18**	**16**
19	**127**	**242**	**19**	**18**
20	**143**	**267**	**20**	**20**

4.B. CORRELATION DE KENDALL



Filet	Cuisse	1	2.5	2.5	8	4	5	13	9	18	15	12	6	7	10	14	11	18	16	18	20	Somme
1	1	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19
2	2.5		-	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
3	2.5			-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
4	8				-	1	-1	1	1	1	1	1	-1	-1	1	1	1	1	1	1	1	8
5	4					-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
6.5	5						-	0	1	1	1	1	1	1	1	1	1	1	1	1	1	13
6.5	13							-	-1	1	1	-1	-1	-1	-1	1	-1	1	1	1	1	1
8.5	9								-	0	1	1	-1	-1	1	1	1	1	1	1	1	7
8.5	18									-	-1	-1	-1	-1	-1	-1	-1	1	-1	1	1	-5
11.5	15										-	0	0	0	-1	-1	-1	1	1	1	1	1
11.5	12											-	0	0	-1	1	-1	1	1	1	1	3
11.5	6												-	0	1	1	1	1	1	1	1	7
11.5	7													-	1	1	1	1	1	1	1	7
14	10														-	1	1	1	1	1	1	6
15	14															-	-1	1	1	1	1	3
16	11																-	1	1	1	1	4
17	18																	-	-1	1	1	1
18	16																		-	1	1	2
19	18																			-	1	1
20	20																				-	0
																				S		127

Y colonne (2.5) < Y ligne (8) ⇨ 1

Y colonne (8) > Y ligne (5) ⇨ -1

Y colonne (2.5) = Y ligne (2.5) ⇨ 0

X colonne (8.5) = X ligne (8.5) ⇨ 0

$$\tau = (2\times127)/(20\times19) = 0.67$$