Master Mentions

Biodiversité, Ecologie, Evolution (B2E)

Gestion de l'Environnement (GE)

Eco-Epidémiologie (Eco-EPI)

HAB711B

Description et inférence

Cours n° 3 & 4

Bastien Mérigot, Maître de Conférences UM
bastien.merigot@umontpellier.fr

Rappel : une **variable** peut principalement être

- **qualitative** :

- binaire : présence –absence d'une espèce

- nominale : le descripteur présente un nombre fini d'états sans caractère ordonné

Par. ex : carnivore, herbivore.

- ordinale : le descripteur présente un nombre fini d'états ordonnés en une séquence logique.

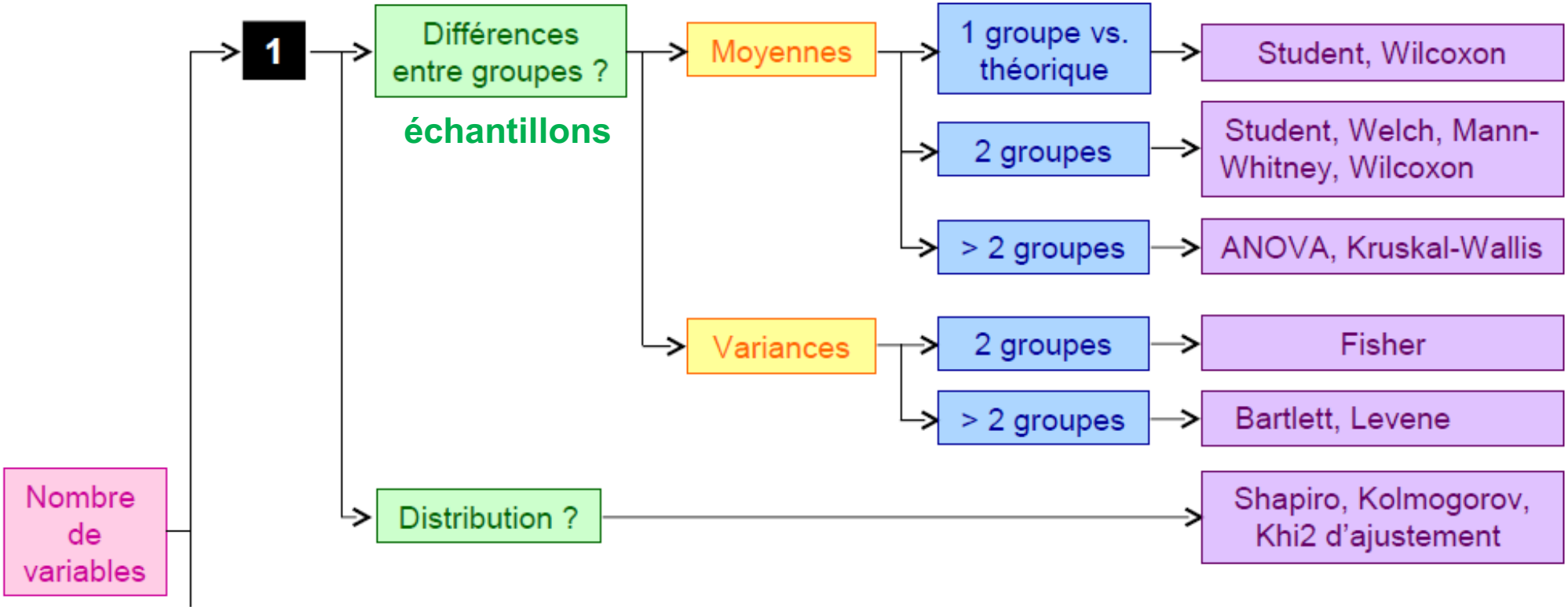
Par. ex : petit, moyen, grand.

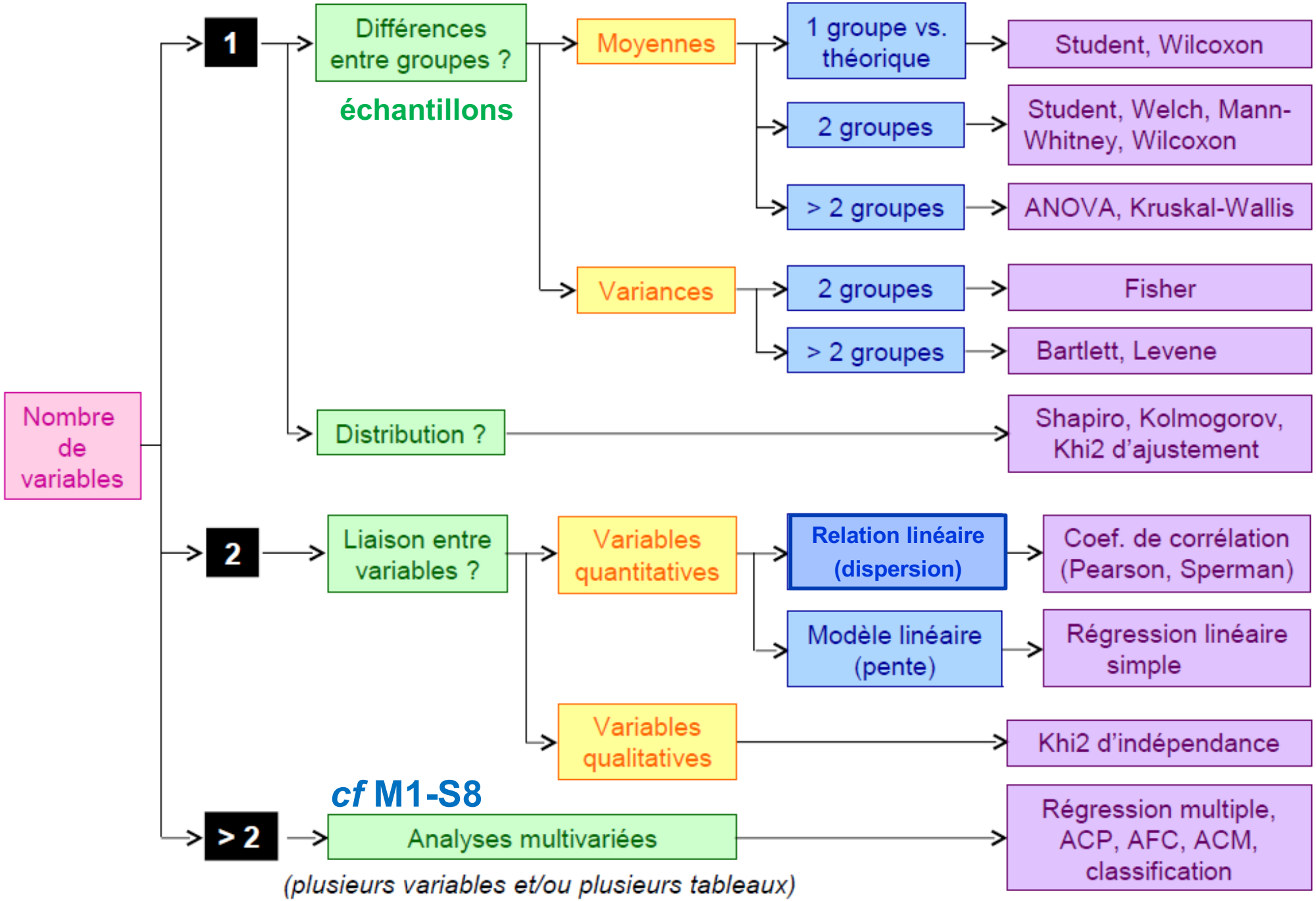
degré de satisfaction mesuré sur une échelle de 0 à 5

- **quantitative** :

- continue : taille, poids

- discrète : nombre d'espèces recensées dans une prairie


Nombre de variables
1
Différences entre groupes ?
échantillons
Moyennes
1 groupe vs. théorique
Student, Wilcoxon
2 groupes
Student, Welch, Mann-Whitney, Wilcoxon
> 2 groupes
ANOVA, Kruskal-Wallis
Variances
2 groupes
Fisher
> 2 groupes
Bartlett, Levene
Distribution ?
Shapiro, Kolmogorov, Khi2 d'ajustement


Nombre de variables
1
Différences entre groupes ?
échantillons
Moyennes
1 groupe vs. théorique
Student, Wilcoxon
2 groupes
Student, Welch, Mann-Whitney, Wilcoxon
> 2 groupes
ANOVA, Kruskal-Wallis
Variances
2 groupes
Fisher
> 2 groupes
Bartlett, Levene
Distribution ?
Shapiro, Kolmogorov, Khi2 d'ajustement
2
Liaison entre variables ?
Variables quantitatives
Relation linéaire (dispersion)
Coef. de corrélation (Pearson, Sperman)
Modèle linéaire (pente)
Régression linéaire simple
Variables qualitatives
Khi2 d'indépendance
cf M1-S8
> 2
Analyses multivariées
Régression multiple, ACP, AFC, ACM, classification
(plusieurs variables et/ou plusieurs tableaux)

Plan CM 3 & 4

I] Test statistique : principe

II] Comparaison de 2 moyennes

Tests de normalité, d'homogénéité des variances,
de comparaison de moyennes (paramétrique) et de rangs (non-paramétrique)

III] Relation entre 2 variables quantitatives

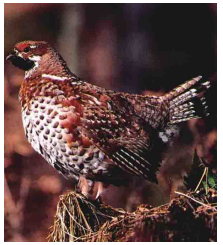
Corrélation, régression linéaire simple

I] Test statistique : principe

Principales étapes d'un test statistique
1) Décrire les données étudiées (type de variables, effectifs, représentations graphiques etc.)
2) Selon la question écologique, poser les hypothèses H0 (nulle) et H1 (alternative)
3) Définir le test approprié pour répondre à la question posée
4) Déterminer si les conditions de validité du test sont remplies et si des tests préalables sont nécessaires à la réalisation du test
5) Choisir le risque de première espèce (alpha)
6) Réaliser le test en calculant la statistique observée et en la comparant à la statistique attendue sous l'hypothèse H0
7) Rejeter ou non l'hypothèse nulle en accord avec le risque alpha
8) Interpréter les résultats au niveau biologique

Diapositives 6 à 21 de F. Leprieur (HMBE101M) actualisées

I. Principe d'un test sur la base d'un exemple : Construction d'un test par permutation



1.Le problème :

On mesure la longueur de la rectrice centrale chez 20 gélinottes huppées mâles juvéniles provenant de 2 sites d'échantillonnage différents. Les résultats sont les suivants :

Site 1	140	150	146	148	152	150	155	153	153	152
Site 2	149	151	152	155	153	157	157	160	158	164

La taille est-elle la même dans la population du site 1 que dans la population du site 2 ?

Taille moyenne pour chacun échantillon :

$$\hat{\mu}_1 = 149.9 \qquad \text{et} \qquad \hat{\mu}_2 = 155.6$$

Est-ce que le fait que la moyenne de taille de l'échantillon 1 est inférieure à la moyenne de taille de l'échantillon 2 nous autorise à dire quelle est aussi plus petite dans la population 1 que dans la population 2 ?

2.Définissons une statistique :

Une mesure qui quantifie la différence entre les 2 longueurs moyennes calculées à partir des échantillons.

Différentes approches :

$$D_1^{obs} = \hat{\mu}_1 - \hat{\mu}_2 = 149.9 - 155.6 = -5.7$$

$$D_2^{obs} = \left|\hat{\mu}_1 - \hat{\mu}_2\right| = \left|149.9 - 155.6\right| = 5.7$$

Si les 2 échantillons sont très différents alors la valeur de la statistique sera grande avec un signe positif ou négative selon le sens dans lequel a été calculé la différence.

Au contraire, si les 2 échantillons proviennent de la même population, la statistique sera petite, proche de zéro, et ne reflètera que des variations aléatoires dues aux fluctuations d'échantillonnage

Les questions posées sont alors :

D_1^{obs}	**est différent de 0 ?**
D_2^{obs}	**est supérieur à 0 ?**

1) Hypothèse que les deux populations sont identiques (hypothèse nulle, H0)

2) Si cela est vrai cela veut dire que nos deux échantillons proviennent de 2 populations identiques en taille de rectrices et donc d'une même population

3) Dans ces conditions, chacun des individus choisis de façon aléatoire pour former les deux échantillons sont des représentants d'une même et unique population.

4) Les longueurs de rectrice observées ne sont que des variations aléatoires autour dune longueur moyenne commune. Si bien que les longueurs observées dans un échantillon auraient tout aussi bien pu être observées dans un autre échantillon.

3) Dans ces conditions, chacun des individus choisis de façon aléatoire pour former les deux échantillons sont des représentants d'une même et unique population.

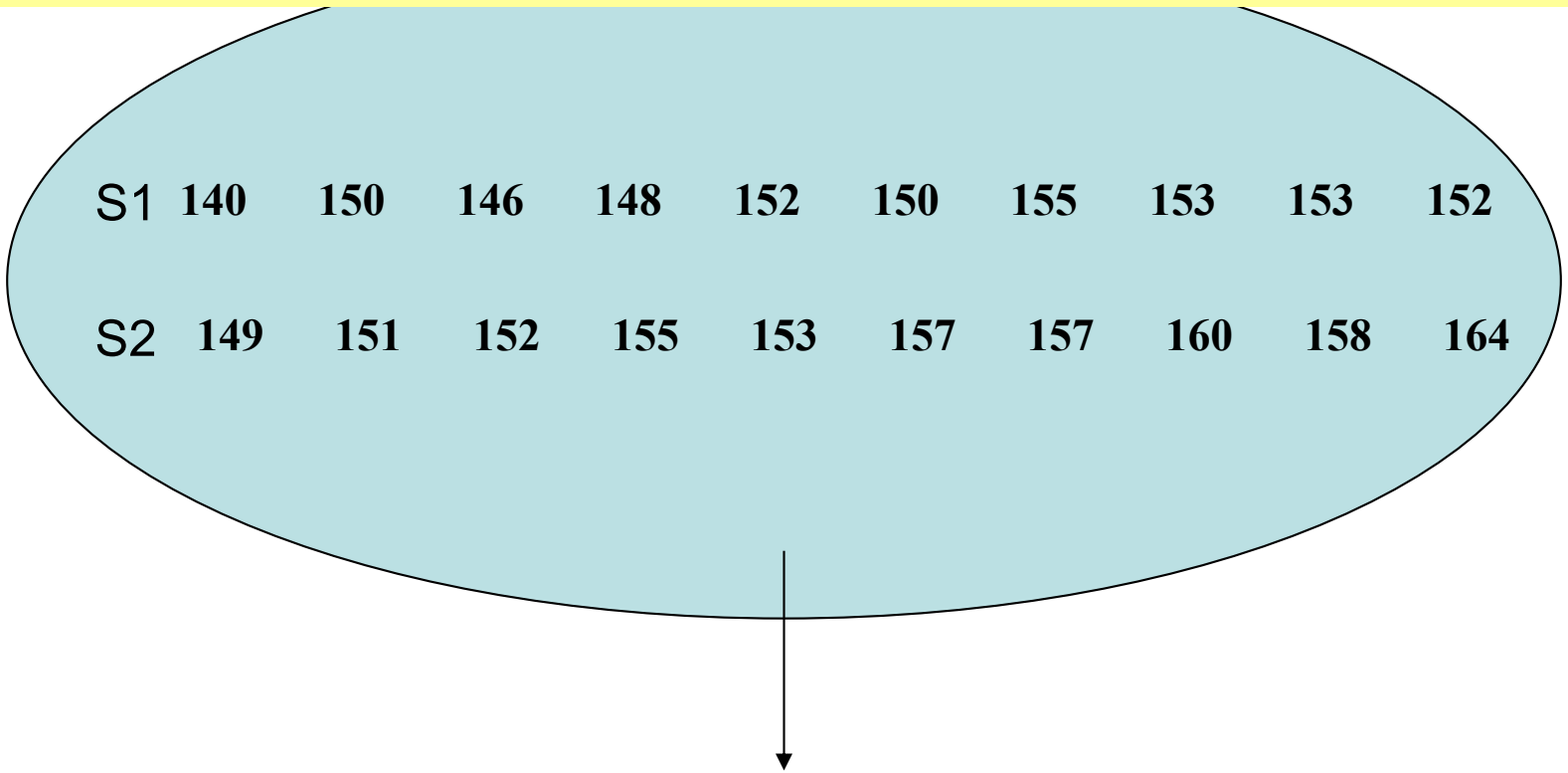
4) Les longueurs de rectrice observées ne sont que des variations aléatoires autour dune longueur moyenne commune. Si bien que les longueurs observées dans un échantillon auraient tout aussi bien pu être observées dans un autre échantillon.

5) Si 10 individus sont choisis au hasard parmi les 20 qui constituent l'ensemble de nos 2 échantillons, on obtient avec ces 10 et les 10 restants, encore 2 échantillons représentatifs de la grande population d'origine.

6) Donc la valeur de la statistique calculée à partir de ces 2 nouveaux échantillons sera une valeur possible de la statistique lorsque l'hypothèse nulle est vraie.

7) En répétant cette opération, les différentes permutations produisent un ensemble de valeurs de la statistique obtenue lorsque l'hypothèse nulle est vraie. L'ensemble de ces valeurs possibles donne une estimation de la distribution d'échantillonnage de la statistique sous Ho.

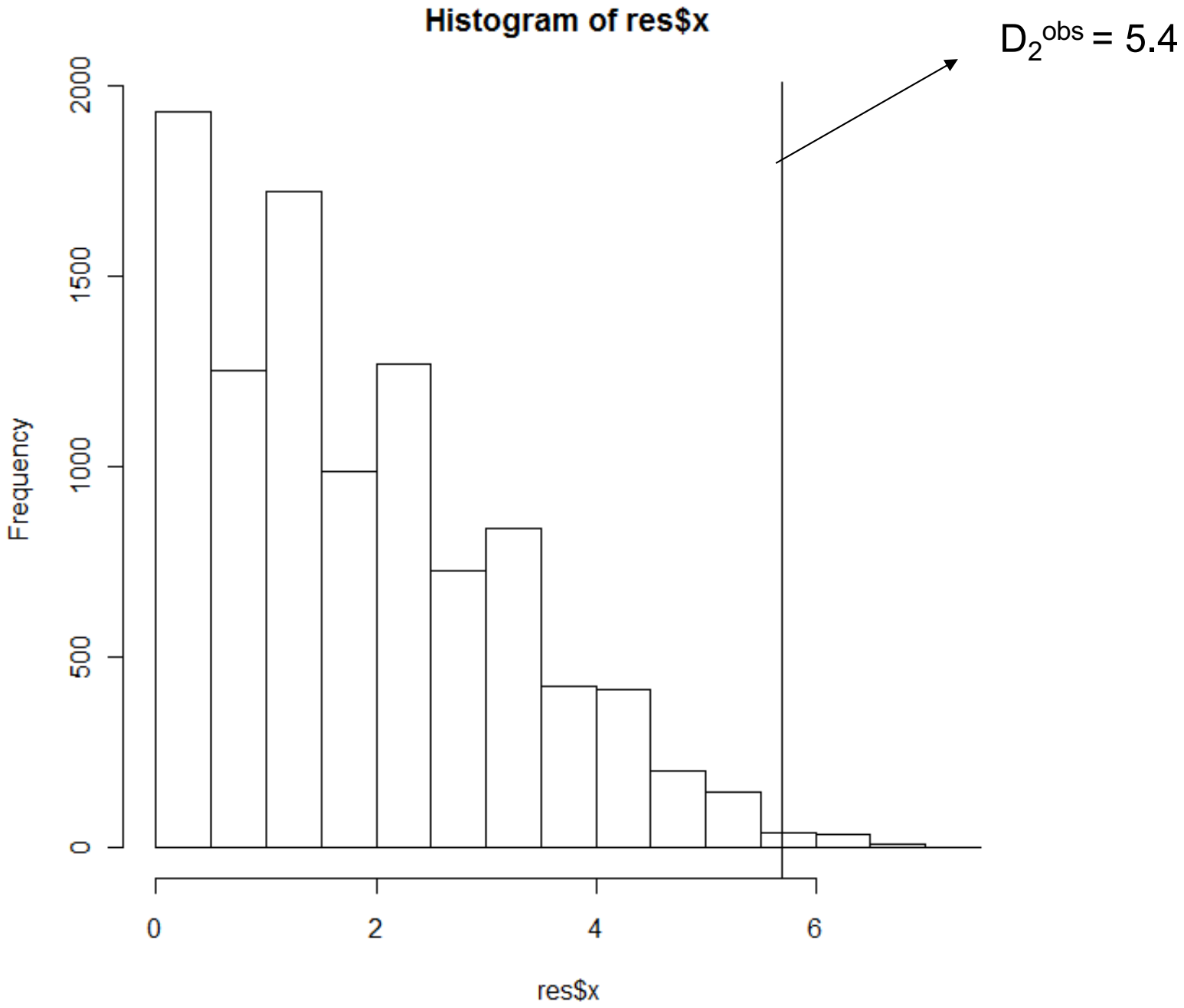
Hypothèse nulle H0 : les deux échantillons proviennent de la même population



On prend au hasard 10 valeurs pour chaque site

S1	140	150	146	148	155	153	155	160	153	164	D1	D2	**Tirage aléatoire 1**
S2	151	150	153	152	157	152	157	158	149	152	-0.70	0.70	
S1	150	157	148	153	150	158	157	149	153	151	D1	D2	**Tirage aléatoire 2**
S2	140	152	153	146	160	152	155	155	152	164	-0.30	0.30	
S1	150	157	160	153	152	152	157	149	164	152	D1	D2	**Tirage aléatoire 3**
S2	140	150	153	146	148	153	151	155	155	158	3.70	3.70	

10 000 tirage aléatoire



Distribution de la statistique D_2 sous l'hypothèse nulle (D_2^{Perm})

Résultats du test

$D_2^{Perm} < D_2^{Obs}$	$D_2^{Perm} = D_2^{Obs}$	$D_2^{Perm} > D_2^{Obs}$
9913	18	69

• Si la valeur observée de la statistique est si élevée qu'elle est plus grande que la plupart des valeurs obtenues en simulant l'hypothèse H0 alors on ne peut pas croire que les résultats expérimentaux sont compatibles avec l'hypothèse nulle et on rejette H0.

• Si au contraire la valeur observée de la statistique se trouve vers le centre de la distribution des valeurs obtenues sous Ho cela montre qu'une telle valeur aurait très bien pu être obtenue au hasard de l'échantillonnage d'une population et on ne rejette donc pas l'hypothèse selon laquelle les données ne sont pas incompatibles avec l'hypothèse nulle aussi appelée hypothèse principale.

La grande majorité des valeurs obtenues sous H0 pour D2 sont inférieures aux valeurs observées de la statistique observée D_2^{obs} (9913 sur 10 000 cas possibles).

Au cours de 10 000 expériences de retirage, seule 69 valeurs de D2 attendues sous l'hypothèse Ho était plus grandes que D_2^{obs} , et 18 valeurs simulées étaient égales à $D_2^{obs.}$

Autrement dit :

Sous l'hypothèse nulle (H0) selon laquelle nos deux échantillons sont issus de populations identiques, la probabilité d'avoir : $D_2^{Perm} \geq D_2^{obs}$

$$P(D_2^{Perm} \geq D_2^{obs} \mid_{H0}) = P(D_2^{Perm} > D_2^{obs} \mid_{H0}) + P(D_2^{Perm} = D_2^{obs} \mid_{H0})$$

$$P(D_2^{Perm} \geq D_2^{obs} \mid_{H0}) = (69+18)/10000 = 0.0087$$

Cette probabilité constitue le risque que l'on prend en rejetant l'hypothèse nulle. C'est le risque de 1ière espèce (erreur de type 1 ou erreur alpha). Ici le risque est faible, donc on peut rejeter H0 en ayant 0.87% de chance de se tromper

0.87% de chance de se tromper en rejetant H0 alors quelle est vraie

Lors d'un test d'hypothèse, on fixe au départ un seuil de risque alpha, le risque que l'on prend en rejetant l'hypothèse nulle.

Si alpha=5% une probabilité est de 0.05 de rejeter H0 alors qu'elle est vraie (pas de différences de longueurs de rectrice centrale chez deux échantillons de Gélinotes) :

- D'après le test, on a obtenu une probabilité de rejeter H0 alors qu'elle est vraie de 0.0087, ce qui est beaucoup plus faible que le seuil de risque alpha que l'on a fixé au début de l'étude.

-On peut donc rejeter H0 au seuil de risque alpha de 5%

- Hypothèse alternative H1 = il existe une différence de moyenne entre les deux échantillons (individus issus de populations différentes)

- Il est possible d'abaisser le seuil de rejet de l'hypothèse nulle (seuil de risque alpha), de façon à n'avancer que des hypothèses très « fiables »

Erreurs de type I et II

L'inconvénient est que, ce faisant, on augmente les chances de commettre une **autre erreur**, celle de ne pas rejeter l'hypothèse nulle alors qu'elle est fausse.

- **erreur de type I** (ou de première espèce) = rejet de l'hypothèse nulle alors qu'elle est vraie

- **erreur de type II** (ou de seconde espèce) = acceptation de l'hypothèse nulle alors qu'elle est fausse.

Deux erreurs antagonistes : abaisser l'une augmente immédiatement l'autre, et la décision que doit prendre le chercheur est un compromis adapté à la situation.

Exemple:

- erreur de type I = condamner un innocent ;
- erreur de type II = laisser un coupable en liberté.

- Avec α la probabilité de l'erreur de type I et β celle de l'erreur de type II, on peut dresser le tableau suivant :

	H_0 vraie	H_0 fausse
acceptation	1 - β	β (erreur II)
rejet	α (erreur I)	1 - α

Comment conclure ?

- Soit on compare la statistique observée (S_{obs}) à la statistique théorique (S_{theo}) attendue sous H0 = utilisation des tables

- Soit on compare la probabilité observée de rejeter H0 à tort au risque alpha fixé au début (ce que fournit les logiciels de statistiques)

Cela revient au même (cf TD et TP)

Deux types de test : test bilatéral vs unilatéral

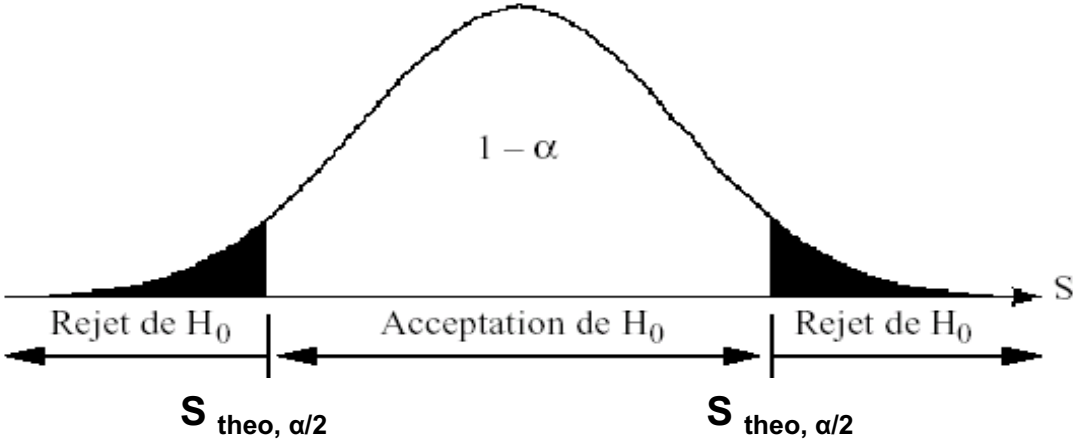
Le choix de l'un ou de l'autre dépend de la question que l'on se pose

1) Il y a-t-il une différence de la taille moyenne de la rectrice centrale entre les deux échantillons (ces échantillons proviennent-ils de deux populations différentes) (Test bilatéral)

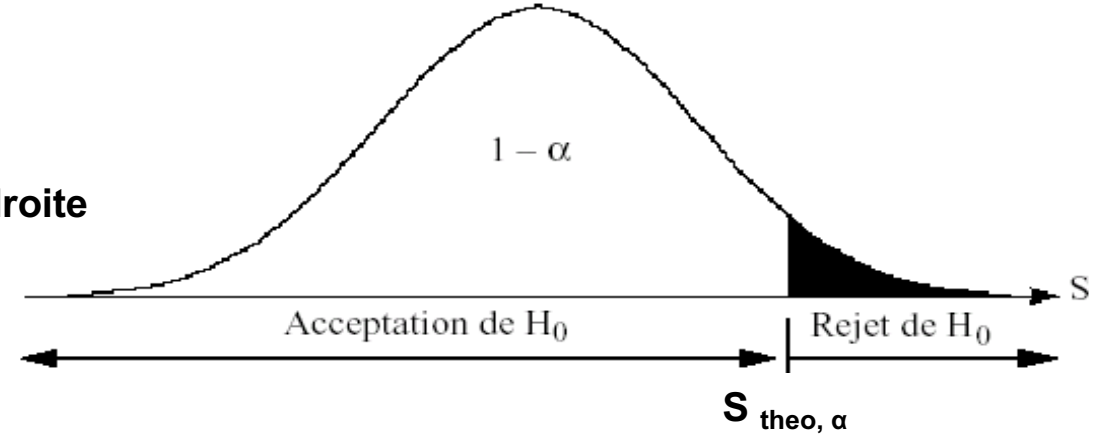
2) Est-ce que la taille moyenne de rectrice centrale de l'échantillon 1 est supérieure à celle de l'échantillon 2 ? (Test unilatéral), et inversement

Distribution de la statistique théorique (attendue sous l'hypothèse nulle : test bilatéral vs. unilatéral)

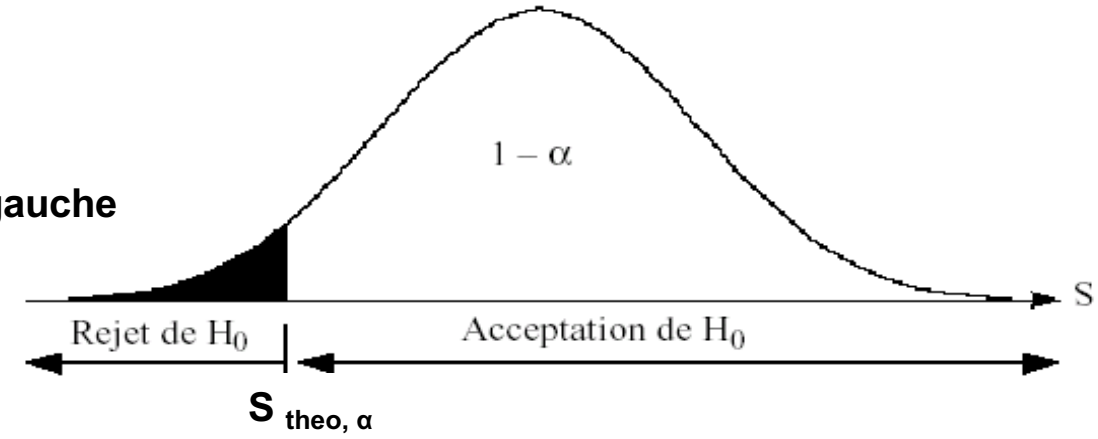
Test bilatéral
H1 : | S | ≥ | $S_{theo,\ \alpha/2}$ |



Test unilatéral à droite
H1 : S ≥ $S_{theo,\ \alpha}$



Test unilatéral à gauche
H1 : S ≤ $S_{theo,\ \alpha}$



Un résultat non statistiquement significatif peut avoir 2 causes :

- l'hypothèse *H0* est vraie (p.ex. il y a équivalence entre les deux échantillons dans le cadre des tests d'égalité)

- la puissance statistique n'est pas suffisante (i.e. nombre d'individus insuffisants)

II] Comparaison de 2 moyennes

Nécessite plusieurs tests successifs :

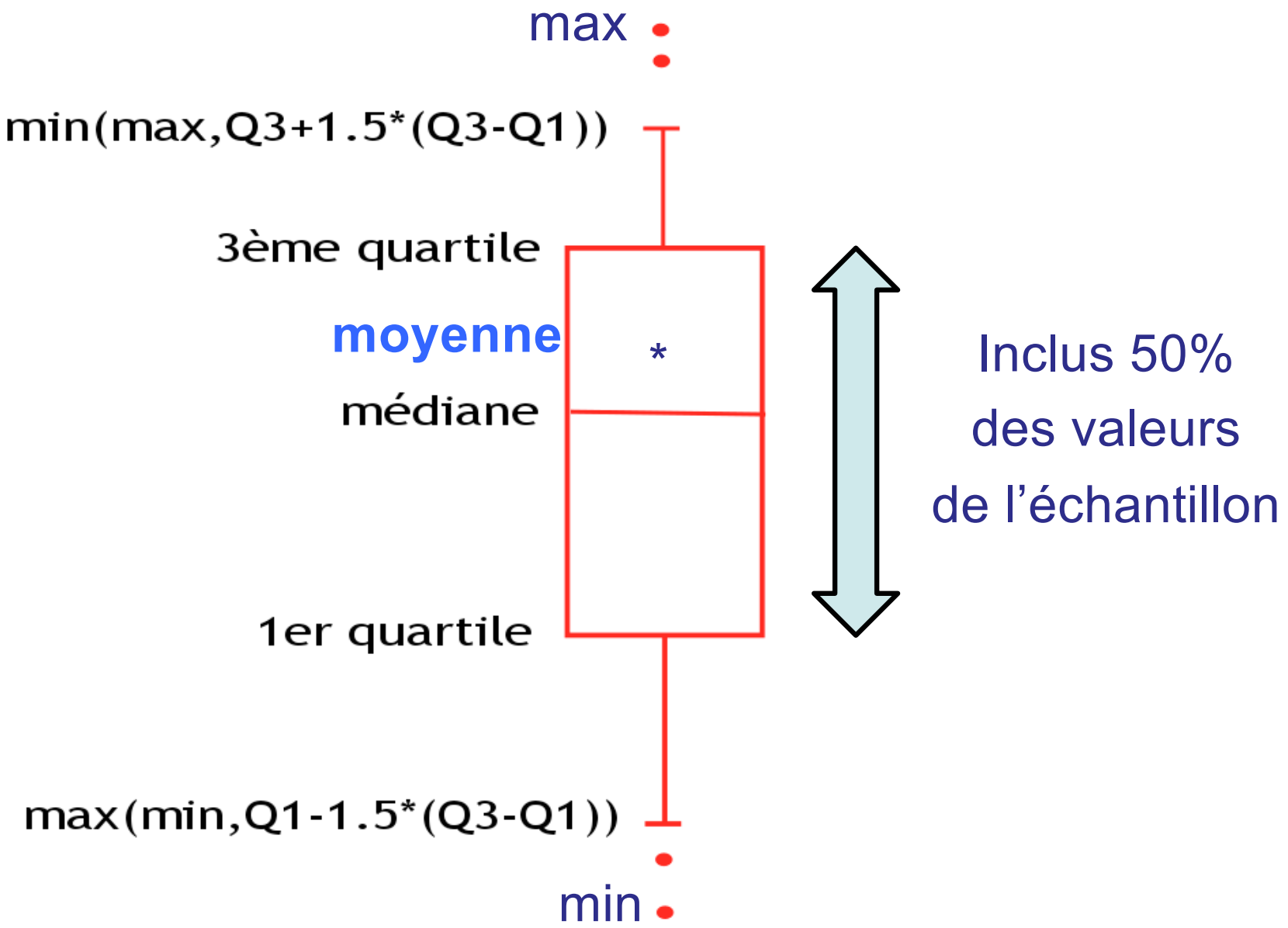
- Test de **normalité** (Shapiro-Wilks)

- Test **d'homogénéité (d'égalité) des variances** (Fisher)

- Test de **comparaison de moyennes** (paramétrique : test t de Student, ou test t avec correction de Welch)

ou de comparaison de distribution de rangs (non-paramétrique : test de Wilcoxon ou de Wilcoxon Mann- Whitney)

Rappel : **Boxplot ou boite à moustaches**



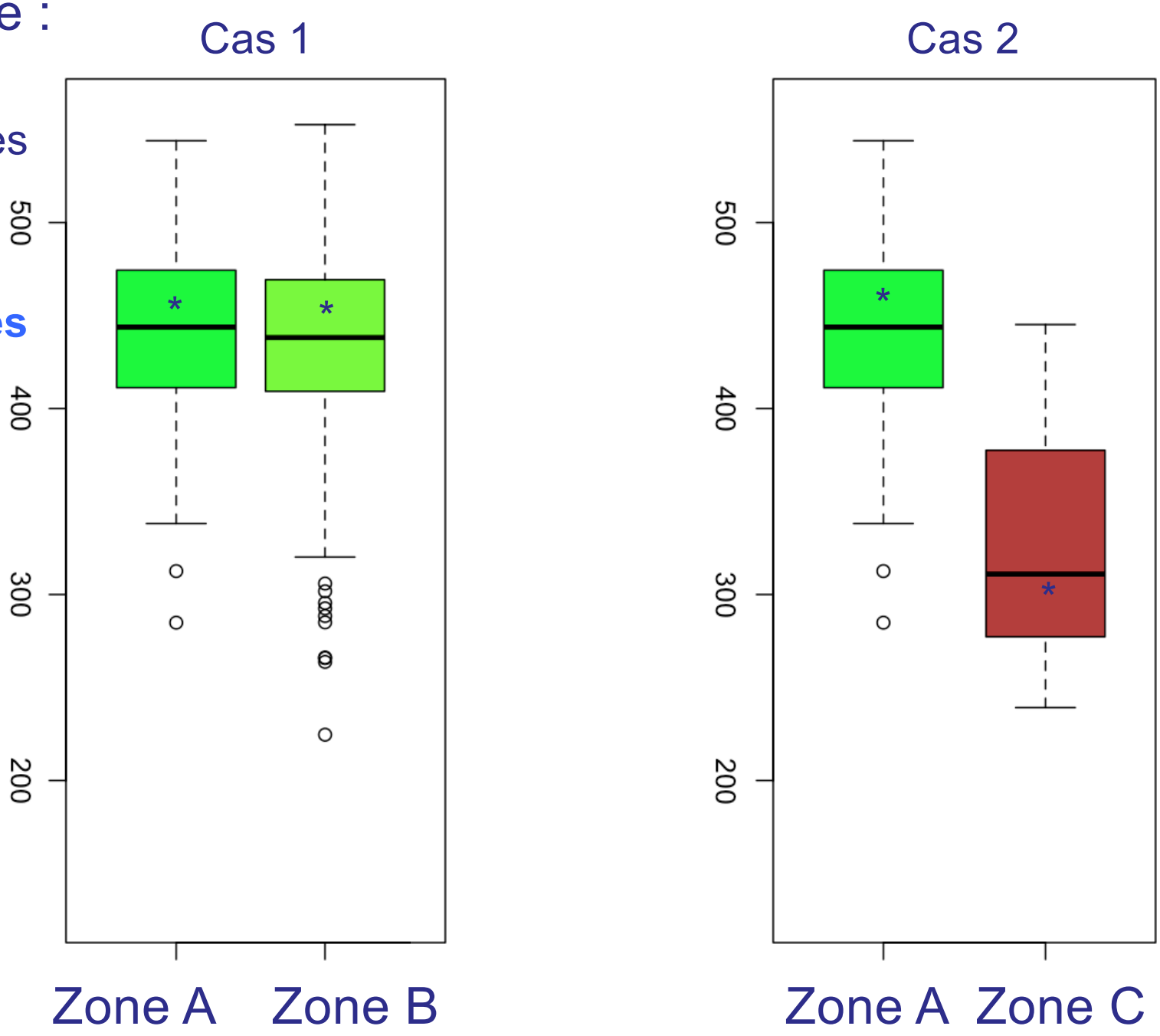
Plusieurs définitions possible des moustaches

1 variable :

Nombre d'espèces

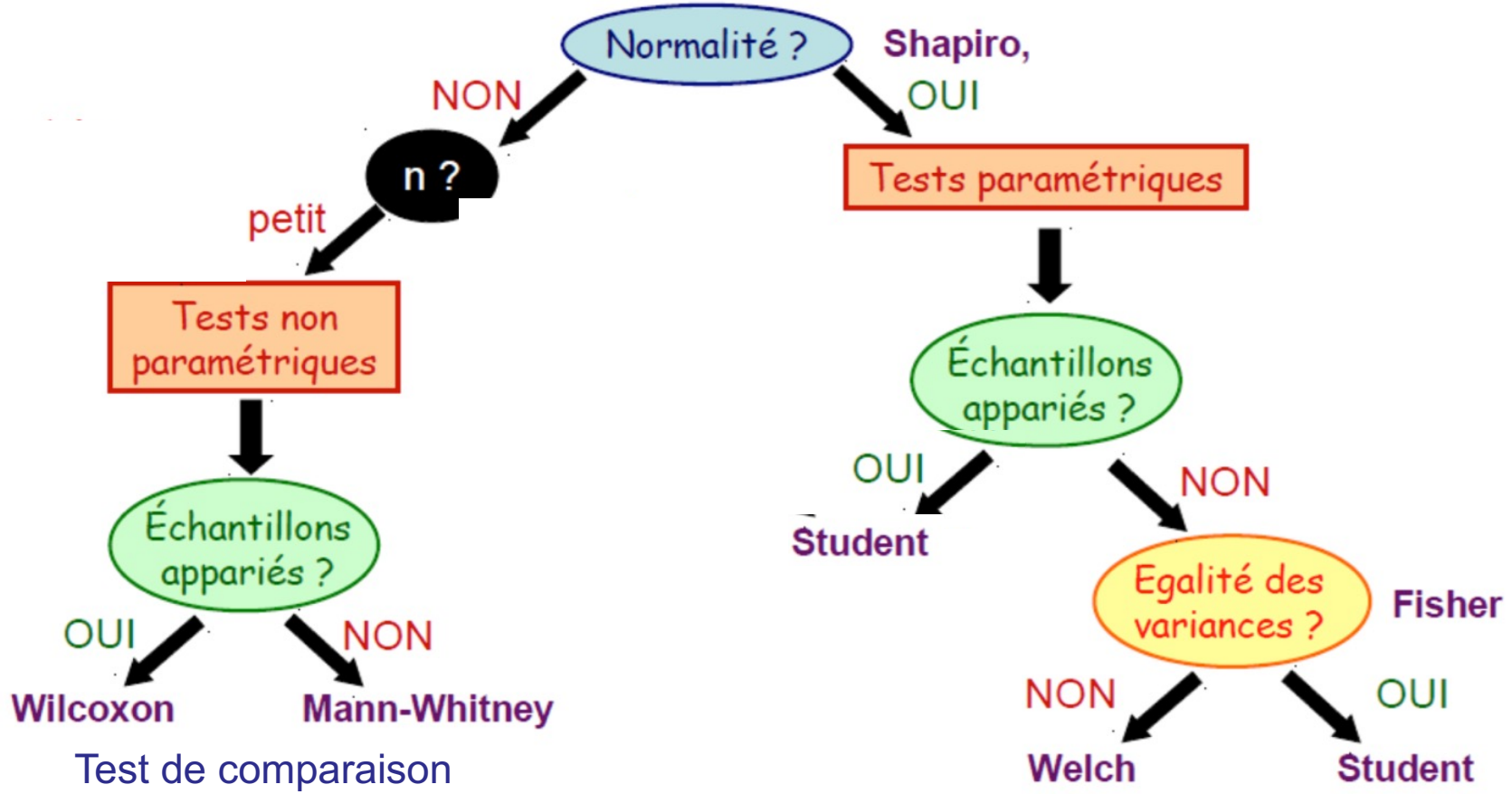
Moyennes

*



n= 30 relevés dans chaque zone

Objectif : comparer les moyennes de 2 échantillons

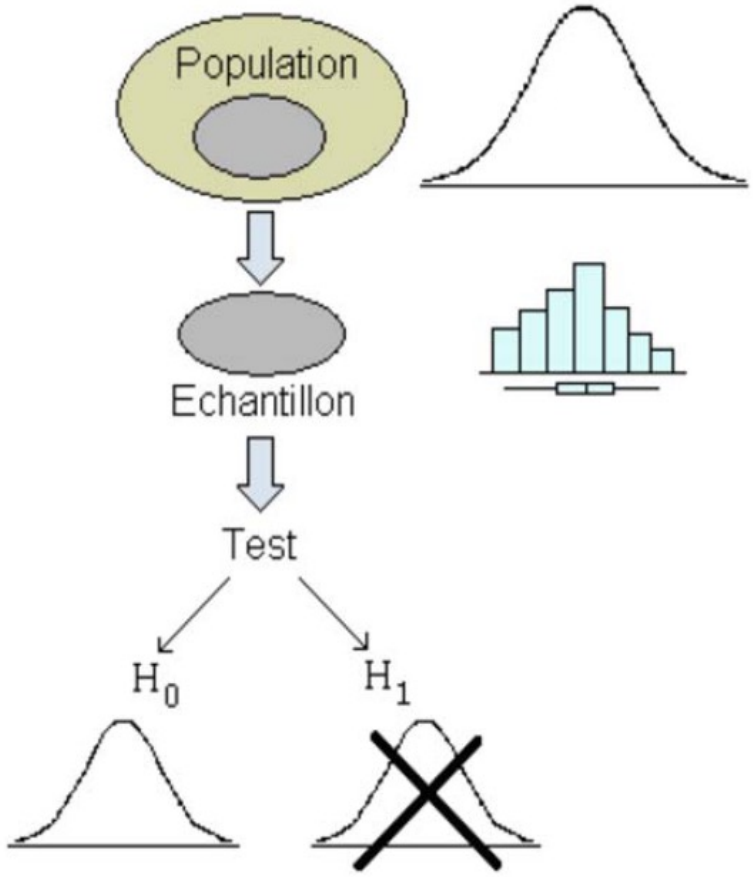


Test de comparaison
des distributions des rangs des valeurs

Test de comparaison
des moyennes (valeurs initiales)

II.1) Test de conformité/ajustement à une loi de distribution

Les tests d'ajustement permettent d'évaluer si un échantillon peut être considéré ou non comme issu d'une loi de distribution donnée.



Pour ce type de test, l'hypothèse nulle spécifie toujours la nature connue/supposée de la distribution de la population

A] Test de normalité de Shapiro-Wilk

Test de normalité basé sur le rapport *W* de deux estimations.

Le rapport *W* sera comparé à une valeur théorique $W_{1-\alpha,n}$ et dans le cas où $W \geq W_{1-\alpha,n}$ nous pourrons accepter, avec un risque d'erreur *α,* l'hypothèse nulle que la distribution suit une loi normale.

Plus *W* est élevé, plus la compatibilité avec la loi normale est crédible.

Procédure

1/ Classement des n valeurs par ordre croissant

$$x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n$$

2/ Calcul de la somme des carrés des écarts (ie. différence par rapport à la moyenne)

$$(1) = SCE = \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2$$

3/ Calcul des différences (étendues partielles)

Si n pair= n/2 différences

Si n impair = (n-1)/2 différences

→ $d_1=x_n-x_1$

$d_2=x_{n-1}-x_2$

$d_3=x_{n-2}-x_3$

4/ Calcul de **b**

Les coefficients a_i sont donnés dans une table

pour n et i donné

→ $$(2) = b = \sum_{i=1}^{n} a_i d_i$$

5/ Calcul de W

→ $$W = \frac{(2)}{(1)} = \frac{b^2}{SCE}$$

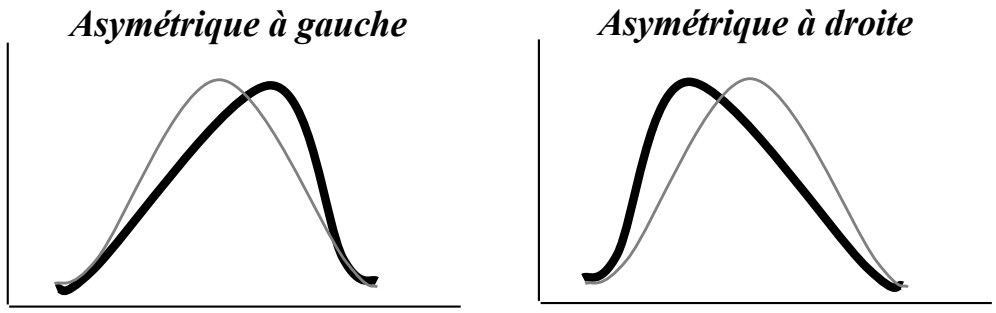
6/ Comparaison de W à $W_{1-\alpha,n}$

$W_{1-\alpha,n}$ dans table de Shapiro-Wilk est fonction de α et n

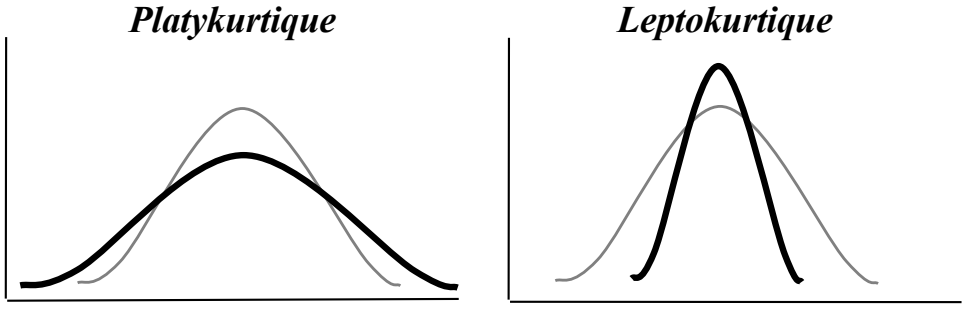
7/ Si $W < W_{1-\alpha,n}$ → H_0 est rejetée → la distribution n'est pas normale

- Sensible aux asymétries
- shapiro.test()

❑ Défaut de symétrie

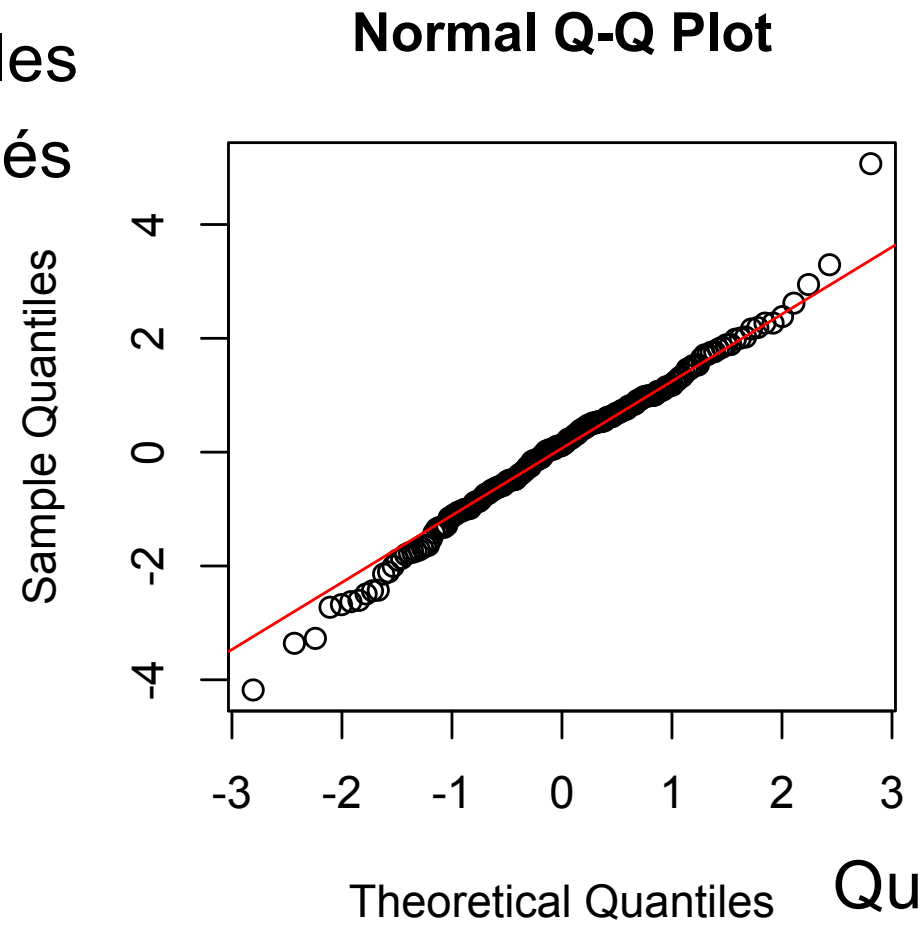


❑ Défaut de variance



Normalité des données : droite de Henry (Tracé Quantile-Quantile)

Quantiles observés



La distribution doit suivre la première bissectrice (ou droite de Henry) d'équation y=x. Dans ce cas distribution normale

Quantiles obtenus pour une loi normale

> qqnorm(y)

> qqline(y, col = 2)

> shapiro.test(y)

W = 0.98869, p-value = 0.1142; Ho : "normalité des données" conservée

Exemple : Mesure de la taille de 15 mammifères marins

Taille (en cm) : 31, 39, 62, 89, 115, 125, 140, 225, 251, 270, 342, 400, 442, 580, 850.

Avec un risque d'erreur de 5%, ces données suivent-elle une loi normale?

Étape 1 : données dans l'ordre croissant :

31, 39, 62, 89, 115, 125, 140, 225, 251, 270, 342, 400, 442, 580, 850

Étape 2 : calcule de la somme des carrés des écarts :

$$
\begin{aligned}
SCE &= \sum (x_i - \overline{x})^2 \\
&= (31 - 264,0\overline{6})^2 + (39 - 264,0\overline{6})^2 + (62 - 264,0\overline{6})^2 + (89 - 264,0\overline{6})^2 \\
&\quad + (115 - 264,0\overline{6})^2 + (125 - 264,0\overline{6})^2 + (140 - 264,0\overline{6})^2 + (225 - 264,0\overline{6})^2 \\
&\quad + (251 - 264,0\overline{6})^2 + (270 - 264,0\overline{6})^2 + (342 - 264,0\overline{6})^2 + (400 - 264,0\overline{6})^2 \\
&\quad + (442 - 264,0\overline{6})^2 + (580 - 264,0\overline{6})^2 + (850 - 264,0\overline{6})^2 \\
&= 734\,482,93
\end{aligned}
$$

Étape 3 : calcul des différences *di*

Différences	d_i
$x_{15} - x_1 =$ 850 - 31	d_1=819
$x_{14} - x_2 =$ 580 - 39	d_2=541
$x_{13} - x_3 =$ 442 - 62	d_3=380
$x_{12} - x_4 =$ 400 - 89	d_4=311
$x_{11} - x_5 =$ 342 - 115	d_5=227
$x_{10} - x_6 =$ 270 - 125	d_6=145
$x_9 - x_7 =$ 251 - 140	d_7=111
$x_8 =$ 225	Aucune : puisque *n* est impair la valeur médiane n'est pas utilisée

Étape 4 : calcul de la valeur de *b*

Pour ce calcul nous avons besoin des coefficients *ai* de la table Shapiro-Wilk pour n = 15.

a_i si n =15	d_i
a_1=0,5150	d_1=819
a_2=0,3306	d_2=541
a_3=0,2495	d_3=380
a_4=0,1878	d_4=311
a_5=0,1353	d_5=227
a_6=0,0880	d_6=145
a_7=0,0433	d_7=111

$$b = \sum a_i d_i = 0{,}5150 \times 819 + 0{,}3306 \times 541 + 0{,}2495 \times 380 + 0{,}1878 \times 311$$
$$+ 0{,}1353 \times 227 + 0{,}0880 \times 145 + 0{,}0433 \times 111 = 802{,}1348$$

Étape 5 : calcul de la statistique *W*

$$W = \frac{b^2}{SCE} = \frac{(802,1348)^2}{734482,93} = 0,876$$

Étape 6 : comparaison de *W* à $W_{1-\alpha,n}$

Avec α=5% et *n = 15 on trouvera dans la table :*

$W_{95\%,15}$*= 0,881*

Puisque 0,876<0,881 on a donc $W < W_{1-\alpha,n}$

la distribution ne suit pas une loi normale avec un risque d'erreur de 5%.

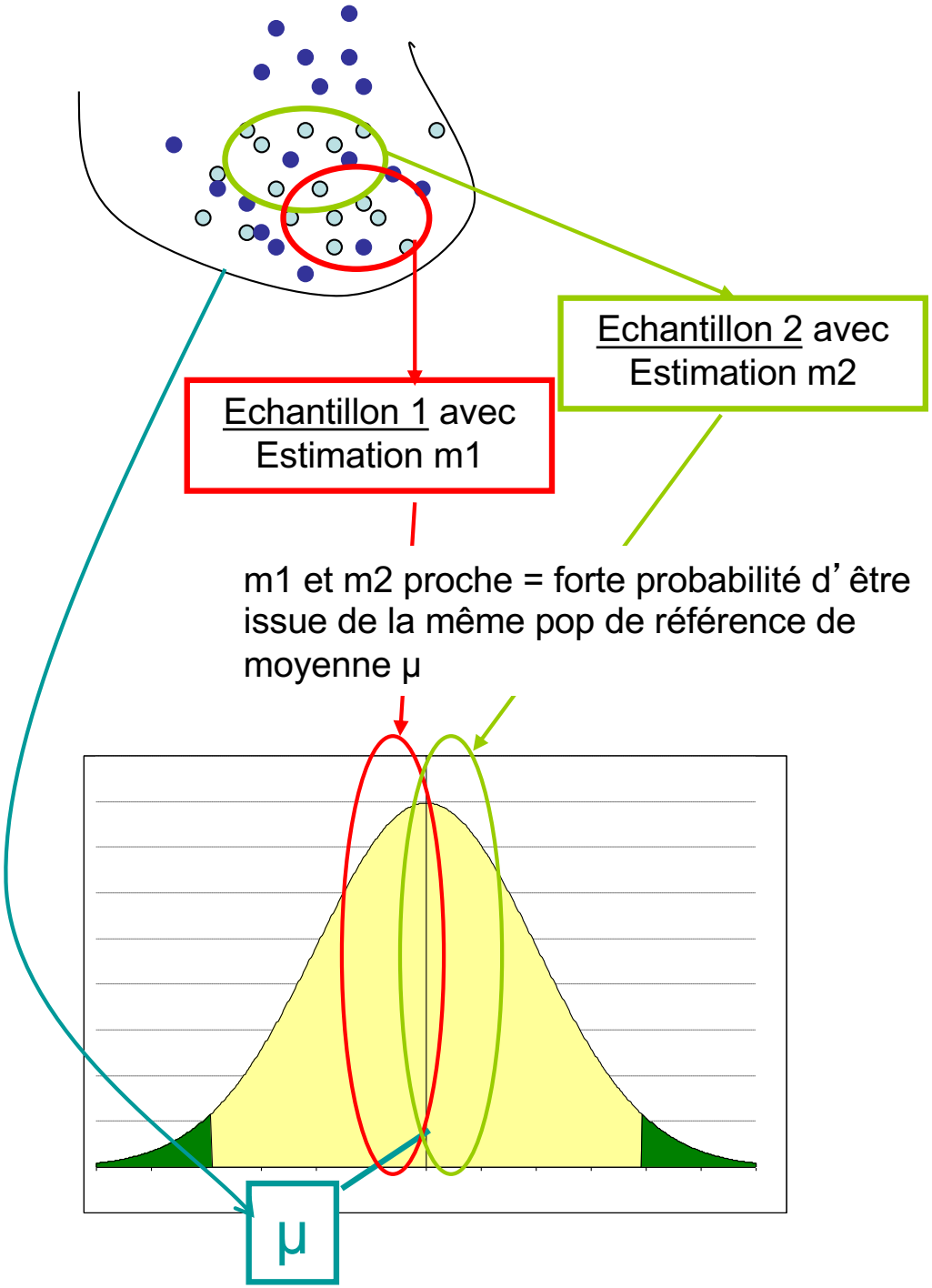
II.2) Tests d'homogénéité ou d'égalité

Est ce que les distributions observées sur deux échantillons sont homogènes et peuvent donc provenir de la même population ?

Problème : On mesure la même quantité dans 2 échantillons, de tailles respectives n_1 et n_2 issus de 2 populations. On cherche à comparer les mesures réalisées dans les 2 populations. Soient X_1 et X_2 les quantités mesurées dans les 2 populations 1 et 2. Les moyennes de X_1 et X_2 sont respectivement μ_1 et μ_2.

Soient
$\hat{\mu}_1$ et $\hat{\mu}_2$ les estimations de ces moyennes réalisées sur les 2 échantillons.

Diapositives 36 à 73 de F. Leprieur (HMBE101M) actualisées


Echantillon 2 avec
Estimation m2
Echantillon 1 avec
Estimation m1
m1 et m2 proche = forte probabilité d'être issue de la même pop de référence de moyenne μ
μ

a) Comparaison de la moyenne de deux échantillons

a-1) approche paramétrique : comparaison de deux variances : le test de student sur échantillons non appariés, et le test de F de test de Fisher-Snedecor

Exemple :

On a mesuré la taille des œufs de coucou en fonction de l'espèce dont le coucou parasite le nid. 2 types de nid ont été échantillonnés, les nids de Pipit des arbres et les nids de l'Accenteur mouchet. Les données suivent sont les suivantes :

X_1	21.05	21.85	22.05	22.45	22.65	23.25	23.25	23.25	23.45	23.45	23.65	23.85	24.05	24.05	24.05

X_2	20.85	21.65	22.05	22.85	23.05	23.05	23.05	23.05	23.45	23.85	23.85	23.85	24.05	25.05	

$n_1 = 15$

$\hat{\mu}_1 = 23.09$

$\hat{\sigma}_1^2 = 0.81$

$n_2 = 14$

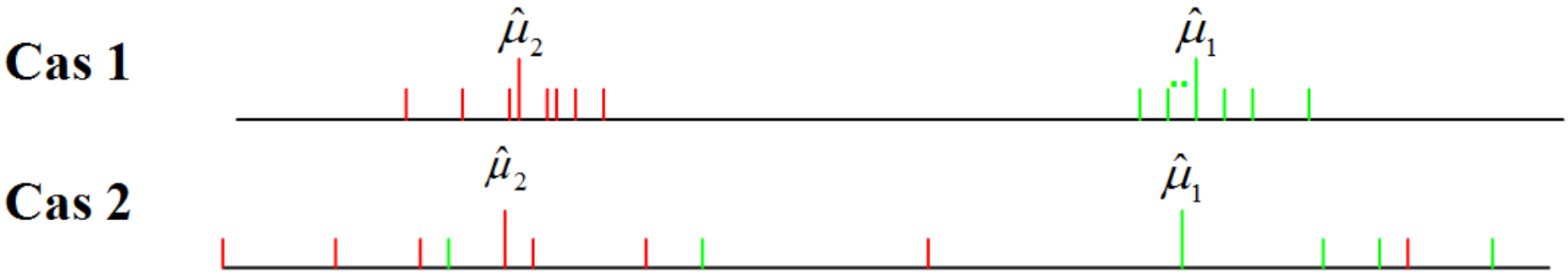
$\hat{\mu}_2 = 23.12$

$\hat{\sigma}_2^2 = 1.142$

On cherche maintenant à tester la différence de taille moyenne entre les œufs de coucou pondus dans les nids de Pipit et ceux pondus dans les nids d'Accenteur.

En d'autres termes, on veut savoir si la différence observée entre les deux moyennes est le fruit de fluctuations fortuites d'échantillonnage ou si elle vient du fait que les échantillons proviennent de 2 populations différentes ?

Graphiquement, nos données sont-elles proches de la distribution du cas 1 ou du cas 2 ?



Pour utiliser le test t de Student pour comparer les moyennes, il faut vérifier avant l'égalité des variances des deux échantillons

Pour vérifier l'homoscédasticté (égalité de variance), le test de *Fisher Snedecor*

H_0: les variances des deux échantillons sont égales : $\sigma^2_1 = \sigma^2_2$
H_1: Les variances des deux échantillons sont différentes : $\sigma^2_1 > \sigma^2_2$ (test unilatéral)

Si les échantillons sont tirés de populations normales (conditions de normalité), le rapport de leurs variances suivra une distribution de *F* à υ**1 = n_1 – 1** et υ**2 = n_2 – 1** degrés de liberté.

Statistique de test $f_{obs} = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$

On rejettera H_0 au seuil de risque alpha si $f_{obs} > F(\alpha, v_1, v_2)$

Attention!!! Il faut **toujours** mettre la **variance la plus grande** au **NUMÉRATEUR**!

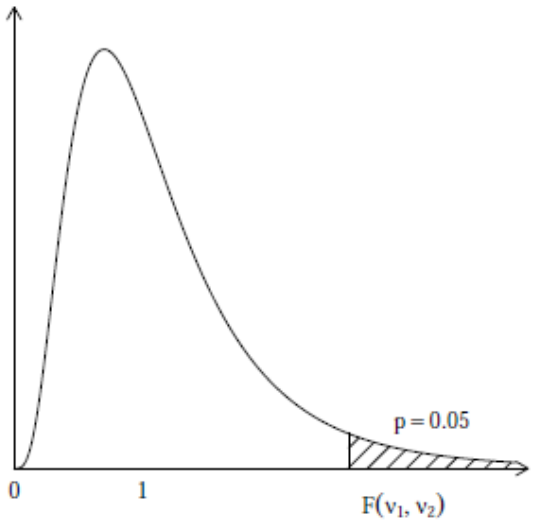


Table de Fisher pour test unilatéral

La statistique de Fisher est définie par

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

avec ici $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$ et où ν_1 est le nombre de d.d.l. correspondant à $\hat{\sigma}_1^2$ et ν_2 celui correspondant à $\hat{\sigma}_2^2$.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	161,4	199,5	215,7	224,6	230,2	234	236,8	238,9	240,5
2	18,51	19	19,16	19,25	19,3	19,33	19,35	19,37	19,38
3	10,13	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999
5	6,608	5,786	5,409	5,192	5,05	4,95	4,876	4,818	4,772
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099
7	5,591	4,737	4,347	4,12	3,972	3,866	3,787	3,726	3,677
8	5,318	4,459	4,066	3,838	3,687	3,581	3,5	3,438	3,388
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,23	3,179
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,02
12	4,747	3,885	3,49	3,259	3,106	2,996	2,913	2,849	2,796
15	4,543	3,682	3,287	3,056	2,901	2,79	2,707	2,641	2,588
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393
24	4,26	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,3
30	4,171	3,316	2,922	2,69	2,534	2,421	2,334	2,266	2,211
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,18	2,124
60	4,001	3,15	2,758	2,525	2,368	2,254	2,167	2,097	2,04
120	3,92	3,072	2,68	2,447	2,29	2,175	2,087	2,016	1,959
∞	3,841	2,996	2,605	2,372	2,214	2,099	2,01	1,938	1,88

ν_2 \ ν_1	9	10	12	15	20	24	30	40	60	120	∞
1	240,5	241,9	243,9	245,9	248	249,1	250,1	251,1	252,2	253,3	254,3
2	19,38	19,4	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,5
3	8,812	8,786	8,745	8,703	8,66	8,639	8,617	8,594	8,572	8,549	8,526
4	5,999	5,964	5,912	5,858	5,803	5,774	5,746	5,717	5,688	5,658	5,628
5	4,772	4,735	4,678	4,619	4,558	4,527	4,496	4,464	4,431	4,398	4,365
6	4,099	4,06	4	3,938	3,874	3,841	3,808	3,774	3,74	3,705	3,669
7	3,677	3,637	3,575	3,511	3,445	3,41	3,376	3,34	3,304	3,267	3,23
8	3,388	3,347	3,284	3,218	3,15	3,115	3,079	3,043	3,005	2,967	2,928
9	3,179	3,137	3,073	3,006	2,936	2,9	2,864	2,826	2,787	2,748	2,707
10	3,02	2,978	2,913	2,845	2,774	2,737	2,7	2,661	2,621	2,58	2,538
12	2,796	2,753	2,687	2,617	2,544	2,505	2,466	2,426	2,384	2,341	2,296
15	2,588	2,544	2,475	2,403	2,328	2,288	2,247	2,204	2,16	2,114	2,066
20	2,393	2,348	2,278	2,203	2,124	2,082	2,039	1,994	1,946	1,896	1,843
24	2,3	2,255	2,183	2,108	2,027	1,984	1,939	1,892	1,842	1,79	1,733
30	2,211	2,165	2,092	2,015	1,932	1,887	1,841	1,792	1,74	1,683	1,622
40	2,124	2,077	2,003	1,924	1,839	1,793	1,744	1,693	1,637	1,577	1,509
60	2,04	1,993	1,917	1,836	1,748	1,7	1,649	1,594	1,534	1,467	1,389
120	1,959	1,91	1,834	1,75	1,659	1,608	1,554	1,495	1,429	1,352	1,254
∞	1,88	1,831	1,752	1,666	1,571	1,517	1,459	1,394	1,318	1,221	1,002

Si les variances sont égales : test t de Student

Sinon test t de Student avec correction de Welch

Test de Student sur échantillons non appariés

Hypothèses :

Ho : $\mu 1 = \mu 2$

H1, Test bilatéral : $\mu 1 \neq \mu 2$

H1, Test unilatéral : $\mu 1 > \mu 2$ ou $\mu 1 < \mu 2$ selon l'a priori sur les données.

Conditions de validité:

1) les variables X1 et X2 sont **indépendantes**
2) Les variables X1 et X2 suivent des **lois normales** de **même variance.** Cette condition (même variance) est à vérifier par un test de comparaison de variance lorsque les variances des populations sont inconnues.

La statistique de student s'écrit (formule 1) :

$$t_{obs} = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

différence de moyennes entre les deux échantillons

Rappel : sous Ho $\hat{\mu}_1 - \hat{\mu}_2 = 0$

Avec la variance commune :

$$\hat{\sigma}^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2} \longrightarrow \hat{\sigma} = \sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}$$

<u>Distribution de la statistique sous Ho :</u> T suit une loi de Student à v = n_1+n_2-2 degrés de liberté. Il y a perte d'un degré de liberté lors du calcul de chacune des deux moyennes $\hat{\mu}_1$ et $\hat{\mu}_2$

Risque de première espèce :

On va alors calculer le risque de première espèce, p ou p-value, associée à t_{obs} que l'on comparera au risque alpha fixé au début de l'étude. On peut comparer aussi directement les statistiques observées et théoriques

Test bilatéral

$$p = P(|T| \geq |t_{obs}| \,|\, Ho)$$

H1: μ1 ≠ μ2

Rejet de Ho si $|t_{obs}| \geq |t_{theo,\,\alpha/2}|$ ou si p < risque alpha

Test unilatéral :

$$p = P(T \geq t_{obs} \,|\, Ho)$$

H1: μ1 > μ2

Rejet de Ho si $t_{obs} \geq t_{theo,\,\alpha}$ ou si p < risque alpha

ou

$$p = P(T \leq t_{obs} \,|\, Ho)$$

H1: μ1 < μ2

Rejet de Ho si $t_{obs} \leq -t_{theo,\,\alpha}$ ou si p < risque alpha

- **Si ce risque est petit, cela voudra dire que l'on peut rejeter Ho avec un risque très faible de se tromper i.e. de rejeter Ho à tort.**
- **Si le risque est grand, on ne pourra pas rejeter Ho.**

Si on montre qu'il y a égalité des variances, on utilise la formule 1

Sinon, on utilise un test de student avec correction de Welch qui implique une autre formule :

$$t_{obs} = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\frac{\hat{\sigma}_1}{n_1} + \frac{\hat{\sigma}_2}{n_2}}}$$

Formule 2

Distribution de la statistique sous Ho : T suit une loi de Student à υ degrés de liberté

$$\upsilon = \frac{\left(\frac{\hat{\sigma}_1}{n_1} + \frac{\hat{\sigma}_2}{n_2}\right)^2}{\frac{\left(\frac{\hat{\sigma}_1}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\hat{\sigma}_2}{n_2}\right)^2}{n_2 - 1}}$$

Le risque de première espèce se calcule de la manière que le test de student sans correction (voir plus haut)

Exemple des coucous

$n_1 = 15$

$\hat{\mu}_1 = 23.09$

$\hat{\sigma}_1^2 = 0.81$

$n_2 = 14$

$\hat{\mu}_2 = 23.12$

$\hat{\sigma}_2^2 = 1.142$

1) On pose les hypothèses

Ho : $\mu 1 = \mu 2$

H1, Test bilatéral : $\mu 1 \neq \mu 2$

H1, Test unilatéral : $\mu 1 > \mu 2$ ou $\mu 1 < \mu 2$ selon l'a priori sur les données.

2) On fait le test de F pour choisir la formule 1 ou la formule 2

H0: $\sigma^2_1 = \sigma^2_2$ alors $f_{obs}=1$
H1: $\sigma^2_1 > \sigma^2_2$ →

f_{obs}=1.142/0.81= 1.409 → On compare à F pour υ1= 14-1= 13 et υ2=15-1=14 et alpha = 0.05

! υ1=degré de liberté correspondant au numérateur (donc c'est la taille de l'échantillon ayant la plus forte variance estimée moins 1) et υ2=degré de liberté correspondant au dénominateur (donc c'est la taille de l'échantillon ayant la plus faible variance estimée moins 1

Table de F

ν_2 \ ν_1	9	10	12	15	20	24	30	40	60	120	∞
1	240,5	241,9	243,9	245,9	248	249,1	250,1	251,1	252,2	253,3	254,3
2	19,38	19,4	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,5
3	8,812	8,786	8,745	8,703	8,66	8,639	8,617	8,594	8,572	8,549	8,526
4	5,999	5,964	5,912	5,858	5,803	5,774	5,746	5,717	5,688	5,658	5,628
5	4,772	4,735	4,678	4,619	4,558	4,527	4,496	4,464	4,431	4,398	4,365
6	4,099	4,06	4	3,938	3,874	3,841	3,808	3,774	3,74	3,705	3,669
7	3,677	3,637	3,575	3,511	3,445	3,41	3,376	3,34	3,304	3,267	3,23
8	3,388	3,347	3,284	3,218	3,15	3,115	3,079	3,043	3,005	2,967	2,928
9	3,179	3,137	3,073	3,006	2,936	2,9	2,864	2,826	2,787	2,748	2,707
10	3,02	2,978	2,913	2,845	2,774	2,737	2,7	2,661	2,621	2,58	2,538
12	2,796	2,753	2,687	2,617	2,544	2,505	2,466	2,426	2,384	2,341	2,296
15	2,588	2,544	2,475	2,403	2,328	2,288	2,247	2,204	2,16	2,114	2,066
20	2,393	2,348	2,278	2,203	2,124	2,082	2,039	1,994	1,946	1,896	1,843
24	2,3	2,255	2,183	2,108	2,027	1,984	1,939	1,892	1,842	1,79	1,733
30	2,211	2,165	2,092	2,015	1,932	1,887	1,841	1,792	1,74	1,683	1,622
40	2,124	2,077	2,003	1,924	1,839	1,793	1,744	1,693	1,637	1,577	1,509
60	2,04	1,993	1,917	1,836	1,748	1,7	1,649	1,594	1,534	1,467	1,389
120	1,959	1,91	1,834	1,75	1,659	1,608	1,554	1,495	1,429	1,352	1,254
∞	1,88	1,831	1,752	1,666	1,571	1,517	1,459	1,394	1,318	1,221	1,002

On encadre $F_{0.05,\ 13,\ 14}$ car on a pas la valeur pour les degrés de liberté 13 et 14

Le f_{theo} se trouve entre 2.403 au minimum et 2.687 au maximum, $f_{obs} < f_{theo}$, on ne rejette pas Ho, on conclut qu'il y a égalité des variances au seuil de risque de 5%. On utilise la formule 1

3) On calcule la statistique observée de student en utilisant la formule 1:

-On calcule la racine carrée de la variance commune (écart type):

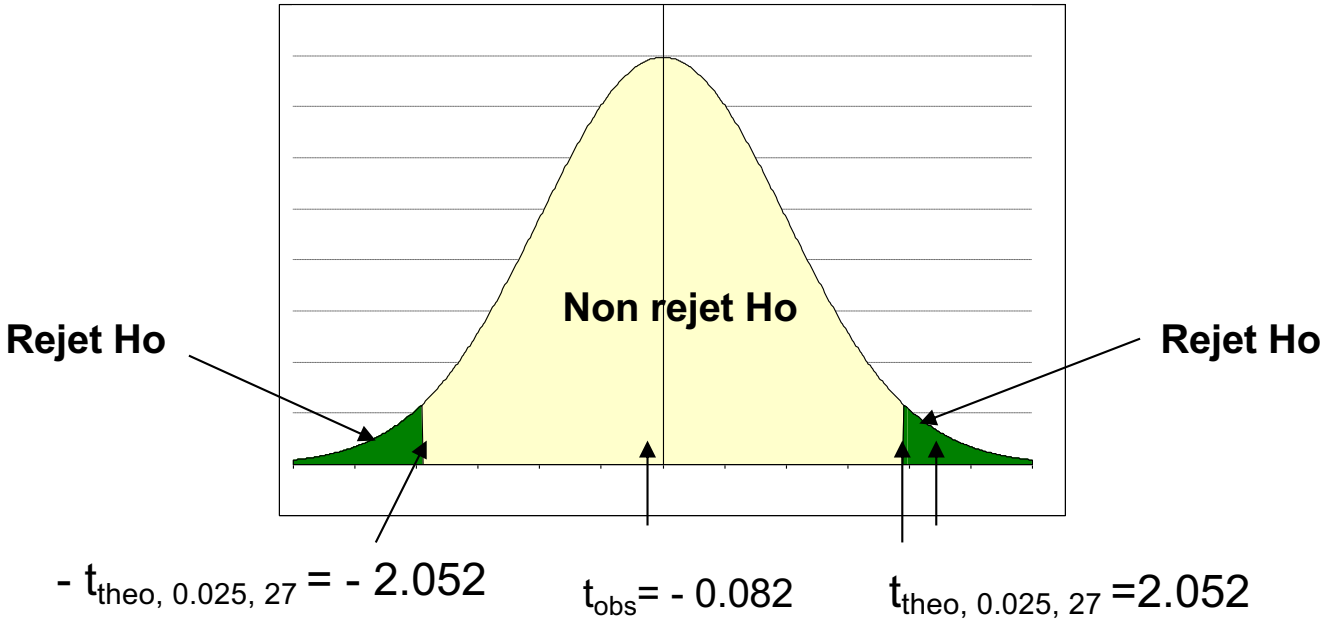
$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}} \qquad \hat{\sigma} = \sqrt{\frac{26.186}{27}} = 0.985$$

- On calcule le t_{obs}

$$t_{obs} = \frac{(23.09 - 23.12)}{0.985\sqrt{\frac{1}{15} + \frac{1}{14}}} = -0.082$$

4) On compare la statistique observée et la statistique théorique de student si H0 est vraie (risque alpha de 5% et υ=15+14-2=27):

On fait ici un test bilatéral : $|t_{theo,\ 0.025,\ 27}|= 2.052$



5) On conclut

$|t_{obs}|< |t_{theo}|$ = non rejet de Ho = pas de différence significative entre les deux moyennes (au seuil de risque alpha de 5%).

On a pas réussi à montrer une différence significative entre les deux échantillons au seuil de risque alpha de 5%. Les 2 moyennes sont différentes

a-2) approche non paramétrique : le test de Wilcoxon-Mann-Whitney (échantillons non appariés) et le test de Wilcoxon (échantillons appariés)

Intérêt des tests non paramétriques :

- **Absence d'hypothèses préalables sur les distributions : normalité, stabilité de la variance. Par contre, l'indépendance des observations demeure une condition requise**

- **Applicables au traitement des échantillons de petite taille,**

- **Basés sur la distributions du rang des valeurs initiales (moins précis), ce ne sont donc pas des tests de comparaisons de moyennes à proprement parlé.**

Conditions de validité: X1 et X2 sont indépendants.

Hypothèse nulle:

H_0: Les rangs des données des deux groupes sont uniformément distribués.

$$\mathbf{P(x_{i,1} > x_{j,2}) = 0,5}$$

Cela signifie que si $x_{i,1}$ est un élément tiré aléatoirement de la première population et $x_{j,2}$ est un élément tiré aléatoirement de la seconde population, il y a une chance sur deux que $x_{i,1}$ soit plus grand que $x_{j,2}$.

Hypothèses alternatives:

- **Test bilatéral**

H_1: Les rangs des données des deux groupes ne sont pas uniformément distribués.

$$\mathbf{P(x_{i,1} > x_{j,2}) \neq 0,5}$$

- **Test unilatéral**

H_1: Les rangs des données du premier groupe sont décalés vers les grandes valeurs.

$$\mathbf{P(x_{i,1} > x_{j,2}) > 0,5}$$

ou H_1: Les rangs des données du premier groupe sont décalés vers les petites valeurs.

$$\mathbf{P(x_{i,1} > x_{j,2}) < 0,5}$$

Ce test ne repose pas sur les paramètres des distributions à l'opposé du test t qui repose sur des paramètres (moyenne, variance)

Le test U de Wilcoxon-Mann-Whitney n'est pas basé sur les valeurs des observations mais sur leurs rangs.

Ce test vise à vérifier si les éléments de deux groupes classés par ordre croissant sur une même échelle ordinale, occupent des positions (rangs) équivalentes révélant ainsi la similitude des deux distributions.

Le degré de mélange des 2 échantillons est mesuré :

On classe l'ensemble des n1 + n2 valeurs mélangées provenant des deux échantillons par ordre croissant et on affecte un rang à chaque valeur (R_{ni}, pour l'individu i de la population 1).

En cas d'ex aequo, on donne un rang moyen à toutes les mesures identiques.

- **Si Ho est vrai, la probabilité d'une valeur de E1 d'être > ou < valeur de E2 =1/2. Les éléments des deux groupes devraient être uniformément mélangés dans ce classement.**

- **Donc Si Ho est vrai, on doit espérer que les rangs moyens des deux échantillons sont égaux.**

- **Si la somme des rangs d'un groupe est très grande ou très petite, on peut supposer que les deux échantillons ne proviennent pas de la même population (hypothèse alternative).**

Exemple 1: deux échantillons A et B

A	B
2	3
4	5
6	7
8	9

Si on classe les données en ordre croissant, ça donne:

A	B	A	B	A	B	A	B
2	3	4	5	6	7	8	9

On voit que les éléments de A et B sont uniformément mélangés: on ne rejetterait probablement pas H_0

Exemple 2: deux échantillons C et D

C	D
2	3
3	6
4	7
5	9

Si on classe les données en ordre croissant, ça donne:

C	C	D	C	C	D	D	D
2	3	3	4	5	6	7	9

On voit que les éléments de D sont en général plus grands: on rejetterait probablement H_0

On calcule alors :

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum_i^n R_{n1}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_1^n R_{n2}$$

où $\sum_i R_{n1}$ et $\sum_i R_{n2}$ **sont les somme des rangs des observations de l'échantillon issu des population 1 et 2 , respectivement.**

Remarque : $U_2 = n_1 n_2 - U_1$ **On peut contrôler que U1 + U2 = n1*n2**

Plus **E1 se mélange à E2** plus la valeurs de U1 et U2 s'approchent de la moyenne de U :

$$\mu_u = \frac{n_1 n_2}{2}$$

Le test consiste à étudier à quel point le plus petit des 2 U (U1 ou U2) dévie de cette valeur moyenne que l'on s'attend à trouver si Ho est vraie.

La statistique de test U_{obs} de Mann-Whitney se définit comme étant la plus petite des deux valeurs U_{n1} et U_{n2} . U_{obs}=min (U_{n1} et U_{n2})

Distribution de la statistique sous H0 et prise de décision:

•**si les échantillons sont grands** (n1 >20 ou n2 > 20) la distribution de U sous Ho est approximativement normale :

de moyenne $\mu_u = \frac{n_1 n_2}{2}$ et de variance $\sigma^2{}_u = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$

On calcule alors $Z_{obs} = \frac{U_{obs} - \mu_u}{\sigma_u}$ qui suit **N(0,1).**

Pour un test bilatéral : si $|Z_{obs}| \geq Z_{theo,\alpha}$, on rejette Ho

Pour un test unilatéral :

1) si H1 : Grp 1< Grp 2, on prend $U_{obs} = U_1$ au lieu de prendre le minimun de U1 et U2

Test unilatéral à gauche : rejet de Ho si $Z_{obs} \leq - Z_{theo, \alpha}$

2) si H1 : Grp 1> Grp 2, on prend $U_{obs} = U_1$ au lieu de prendre le minimun de U1 et U2

Test unilatéral à droite : rejet de Ho si $Z_{obs} \geq Z_{theo, \alpha}$

= Il suffit de confronter Z_{obs} aux valeurs théoriques de la loi normale centrée réduite

- Pour un test bilatéral : si | Zobs | ≥ $Z_{theo,\alpha}$, on rejette Ho

- Pour un test unilatéral :

1) si H1 : Grp 1< Grp 2, on prend $U_{obs} = U_1$

Test unilatéral à gauche : rejet de Ho si $Z_{obs} \leq - Z_{theo, \alpha}$

2) si H1 : Grp 1> Grp 2, on prend $U_{obs} = U_1$

Test unilatéral à droite : rejet de Ho si $Z_{obs} \geq Z_{theo, \alpha}$

Si plusieurs éléments occupent le même rang, une formule corrigée de σ_U doit être utilisée:

$$\sigma_U = \sqrt{\frac{n_1 \times n_2}{n \times (n-1)} \times \left(\frac{n^3 - n}{12} - \sum_{l=1}^{g} E_l \right)}$$

où $n = n_1 + n_2$

g = nombre de rangs avec données ex-aequo

l = le l-ième rang avec ex-aequo

$E_l = \frac{e_l^3 - e_l}{12}$ où e_l = nombre d'observations de rang l

C' est généralement implémenté dans les logiciels de statistique

<u>Si n1 ou n2 < 20,</u> on ne peut pas utiliser la loi normale centrée réduite et on utilise la véritable distribution de U qui est connue. La table de U existe pour n1 et n2 ≤ 20

La table de U que l'on vous a distribué comporte les U_{theo} correspondant à la probabilité $P(U_{obs} \leq U_{theo}) = 0.05$ et 0.01

Pour un test bilatéral :

Si $U_{obs} \leq U_{theo}$ on rejete Ho (c'est le contraire pour les test paramétriques et la table de student attention!)

Pour un test unilatéral :

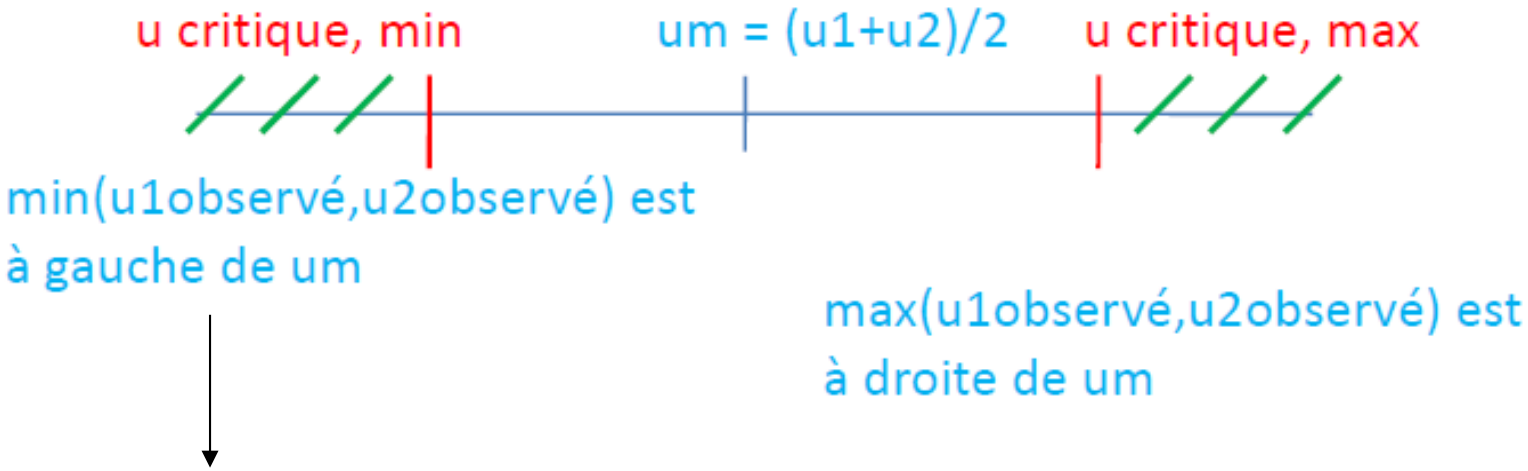
si H1 : Grp 1< Grp 2, on prend $U_{obs} = U_2$ **Rejet de Ho si Uobs ≤ Utheo**

si H1 : Grp 1> Grp 2, on prend $U_{obs} = U_1$ **Rejet de Ho si Uobs ≤ Utheo**

Comme U_{obs}, on prend le U parmi les deux calculés (U_1 ou U_2) qui correspond à l'échantillon dont on fait l'hypothèse qu'il comporte des valeurs plus grandes que l'autre échantillon

Test de Wilcoxon-Mann-Whitney pour des données indépendantes

La table donne seulement les valeurs max de la valeur critique



On vous fournit cette table!

NB: Les valeurs en bleu sont observées, les valeurs en rouge sont fournies par la table correspondant au test

Si les valeurs observées sont dans la partie hachurée en vert à gauche ou à droite, vous devez rejeter l'hypothèse nulle.

Table des valeurs minimales de la statistique théorique U (test bilatéral)

n_2	α	n_1 = 3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54

$$U_{0,05,8,9} = 15$$

Puissance du test *U*

- **Lorsque les conditions d'application du test *t* sont remplies, la puissance du test *U* est près de 95% de celle du test *t*. Les deux tests ont donc approximativement la même capacité de détecter de petites différences entre des échantillons tirés de deux populations normales qui sont statistiquement différentes. Dans certaines autres conditions, le test *U* est plus puissant que le test *t*.**

- **Pour des données tirées de populations à distribution fortement non normale, le test *t* est trop conservateur. Par contre, le taux d'erreur de type I du test *U* est approximativement égal au seuil de signification a dans ces conditions. On emploiera donc ce test.**

Exemple : 2 échantillons C et D

C	D
2	2
3	4
5	5
5	6
7	6
8	7
9	8
10	11
	12

Calcul de *U* :

Rang	Éléments	C	D
1	2	1,5	
2	2		1,5
3	3	3	
4	4		4
5	5	6	
6	5	6	
7	5		6
8	6		8,5
9	6		8,5
10	7	10,5	
11	7		10,5
12	8	12,5	
13	8		12,5
14	9	14	
15	10	15	
16	11		16
17	12		17
	Sommes	$R_1 = 68,5$	$R_2 = 84,5$

$$U_1 = n_1 \times n_2 + \frac{n_1 \times (n_1 + 1)}{2} - R_1 = 8 \times 9 + \frac{8 \times (8+1)}{2} - 68,5 = 39,5$$

$$U_2 = n_1 \times n_2 + \frac{n_2 \times (n_2 + 1)}{2} - R_2 = 8 \times 9 + \frac{9 \times (9+1)}{2} - 84,5 = 32,5$$

Test bilatéral : U_{obs} = min(U_1, U_2) = 32,5

On compare le U_{obs} au $U_{théorique}$

Si la **valeur de U_{obs}** à partir des données est **plus petite** que la valeur de la table, on **rejette H_0.**

U_{obs} = 32,5
n_C = 8
n_D = 9
α = 0,05 bilatéral

Dans la table, la valeur de U_{theo} pour un seuil α = 0,05 bilatéral est $\boldsymbol{U_{0,05,8,9}}$ **= 15**

Puisque le U_{obs}=32,5, et 32,5 > 15 on ne rejette pas H_0

Autre exemple

Longueur des ailes antérieures droites (mm) des mâles de *Papilio glaucus* échantillonnés en Alaska et en Illinois une certaine année. On considère ici que la longueur des ailes ne suivent pas une loi normale.

Alaska	Illinois
42	51
41	48
41	49
37	48
44	47
43	46
43	47
40	47
40	50
44	54

1) On pose les hypothèses

H0: Les rangs des données des deux groupes sont uniformément distribués. La taille des ailles de papillons en Alaska est identique à celle des papillons d'Illinois

H1 : Test bilatéral: Les rangs des données des deux groupes ne sont pas uniformément distribués. La taille des ailes de papillons en Alaska est différente de celles des papillons d'Illinois

H1 : Test unilatéral: Les rangs des données du premier groupe sont décalés vers les petites valeurs. La taille des ailes de papillons en Alaska est plus petite que celles des papillons d'Illinois

2) On calcule la statistique observée U_{obs}

Ordre croissant	Groupes	Longueur	Rangs
1	A	37	1
2	A	40	2.5
3	A	40	2.5
4	A	41	4.5
5	A	41	4.5
6	A	42	6
7	A	43	7.5
8	A	43	7.5
9	A	44	9.5
10	A	44	9.5
11	I	46	11
12	I	47	13
13	I	47	13
14	I	47	13
15	I	48	15.5
16	I	48	15.5
17	I	49	17
18	I	50	18
19	I	51	19
20	I	54	20

n1=10 n2=10

$$\sum_i R_A = 55 \qquad \sum_i R_I = 155$$

$$U_1 = 10 \times 10 + \frac{10 \times (10+1)}{2} - 55$$

$$U_1 = 155 - 55 = 100$$

$$U_2 = 10 \times 10 + \frac{10 \times (10+1)}{2} - 155$$

$$U_2 = 155 - 155 = 0$$

Pour un test bilatéral U_{obs}= min (U1, U2) = 0

3) On compare la statistique observée à la statistique théorique

On utilise la table de Mann-Whitney car on a des échantillons de petite taille (<20)

1) Test bilatéral : On cherche U_{theo} pour n_1=10 et n_2=10 et on fixe un seuil de risque alpha de 5%

On trouve $U_{theo, 10,10, 0.05}$= 23

U_{obs}<U_{theo} donc on rejette Ho : la taille des ailes de papillons d'Alaska est significativement différente de celle des papillons d'Illinois au seuil de risque alpha de 5 %

2) Test unilatéral : On cherche U_{theo} pour n_1=10 et n_2=10 et on fixe un seuil de risque alpha de 5%

On trouve $U_{theo, 10,10, 0.05}$= 27

Ici on cherche à tester l'hypothèse H1 : Alaska<Illinois , on prend donc U_{obs}= U_2 c'est-à-dire $U_{illinois}$

U_{obs}=U_2= 0 <U_{theo} donc on rejette Ho : la taille des ailes de papillons d'Illinois est significativement plus grande que celle des papillons d'Alaska au seuil de risque alpha de 5 %

2) Test unilatéral : On cherche U_{theo} pour n1=10 et n2=10 et on fixe un seuil de risque alpha de 5%

On trouve $U_{theo, 10,10, 0.05}$= 27

Ici on cherche à tester l'hypothèse H1 : Alaska>Illinois , on prend donc U_{obs}= U_1 c'est-à-dire U_{Alaska}

U_{obs}=U_1= 55 > U_{theo} donc on ne rejette Ho : la taille des ailes de papillons d'Alaska n'est pas significativement plus grande que celle des papillons d'Illinois au seuil de risque alpha de 5 %

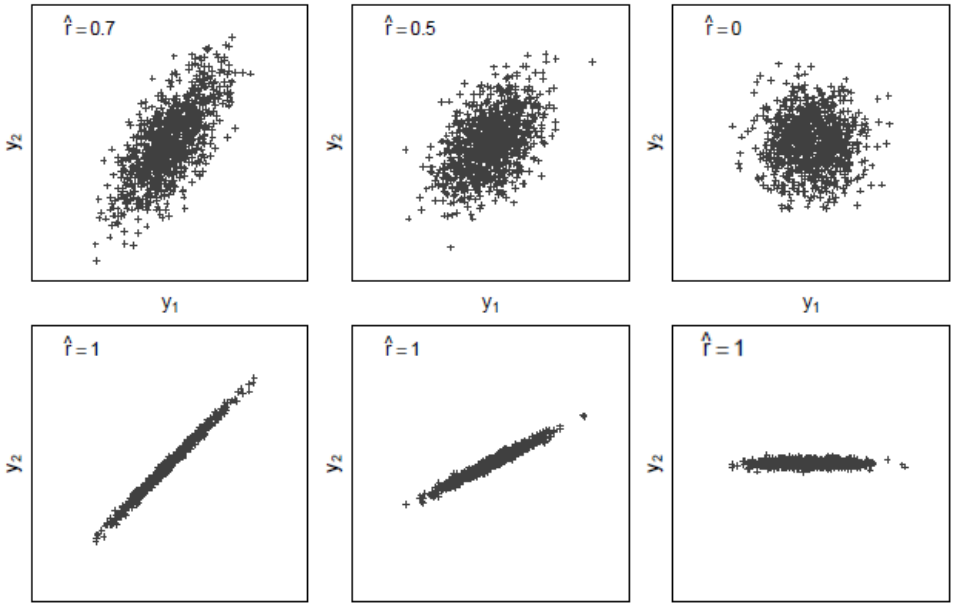
III] Relation entre 2 variables quantitatives

Corrélation, régression linéaire simple

1) Correlations de Pearson et Spearman et tests associés

a) Coefficient de corrélation de Pearson.

Le coefficient de corrélation linéaire mesure la dispersion autour de la relation moyenne entre 2 variables quantitatives



≠ avec la pente *a* qui mesure l'intensité de la relation linéaire

On observe deux caractères sur un même individu

$$X=\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} \qquad Y=\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

(x_i,y_i) est le couple d'observations mesurées sur l'individu i

n est taille de l'échantillon

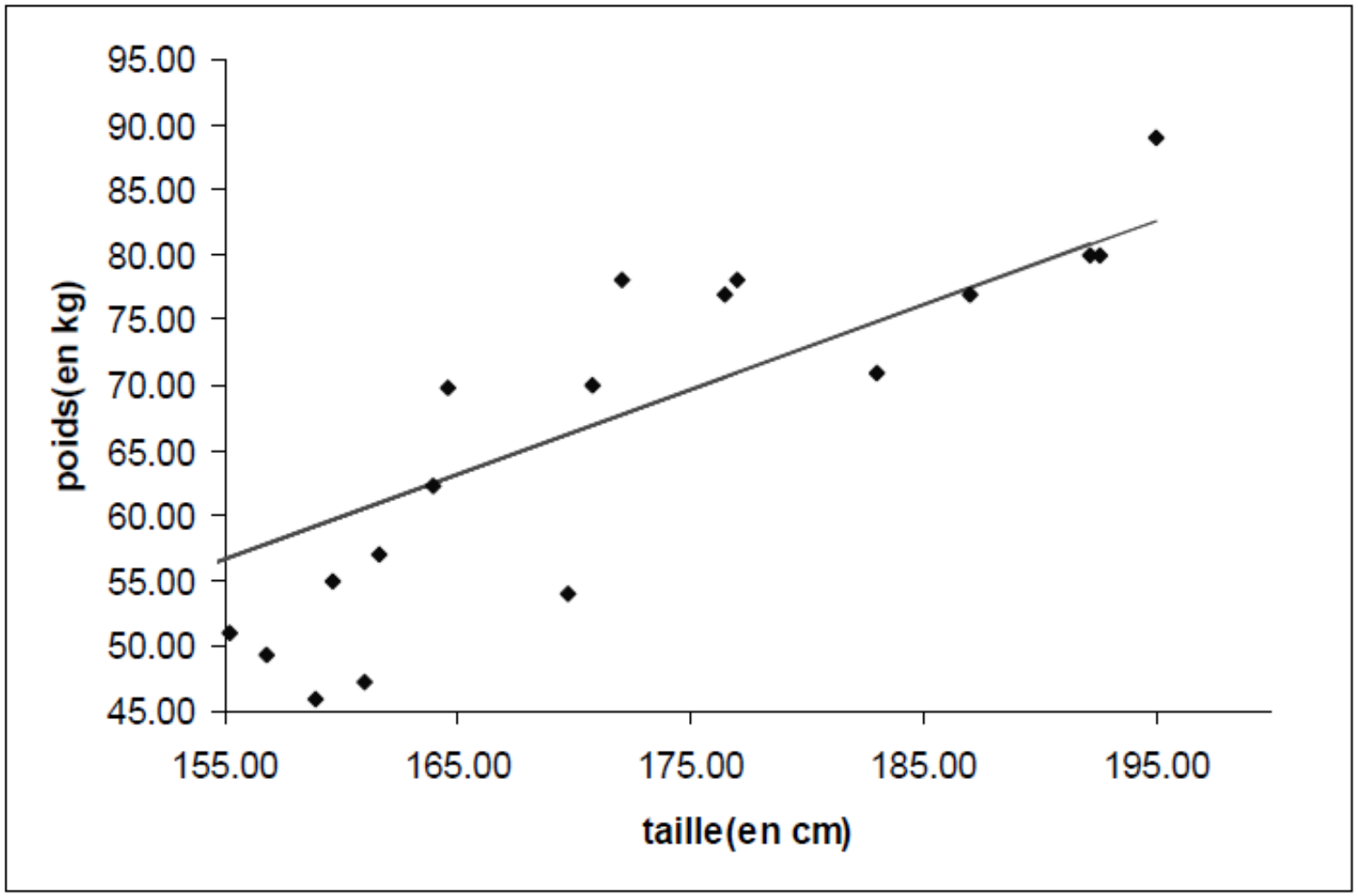
X : taille (cm)

Y : poids (kg)

n =10

X	Y
174.0	70.4
172.1	57.5
159.8	50.0
173.9	69.9
162.9	62.1
174.1	67.3
178.7	78.0
161.6	57.5
180.2	76.9
170.7	63.3

Un point correspond à 1 individu pour lequel 2 variables ont été mesurés (poids et taille)



On dit qu'il y a corrélation entre deux variables X et Y si il y a dépendance en moyenne

à X=x fixé la moyenne des y_i est fonction de x

• Contexte : analyse simultanée de 2 variables quantitatives mesurées sur les mêmes individus → comment co-varient elles ?

• Covariance : mesure de la dispersion conjointe de 2 variables autour de leur moyenne

Variance

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Covariance

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

Valeur de covariance dépend des unités de mesure des variables

Coefficient de corrélation de Pearson *r* standardise la covariance :

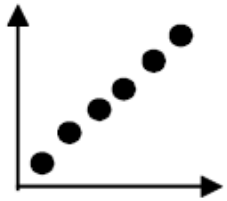
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \times \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Covariance de X et Y

Produit des écarts types de X et Y

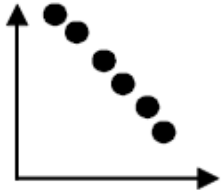
- $-1 \leq r \leq 1$
- Si X et Y sont indépendants alors : $|r| = 0$
- Si il existe une corrélation linéaire alors : $|r| \approx 1$

. La corrélation est importante si I r I > 0.6 ou 0.7 (subjectif)



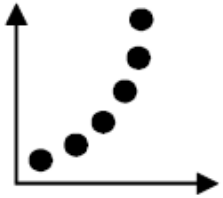
A
corrélation positive

liaison : **monotone linéaire croissante**



B
corrélation négative

monotone linéaire décroissante

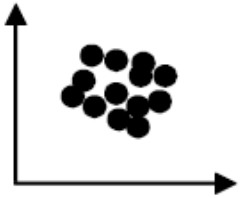


C
corrélation positive

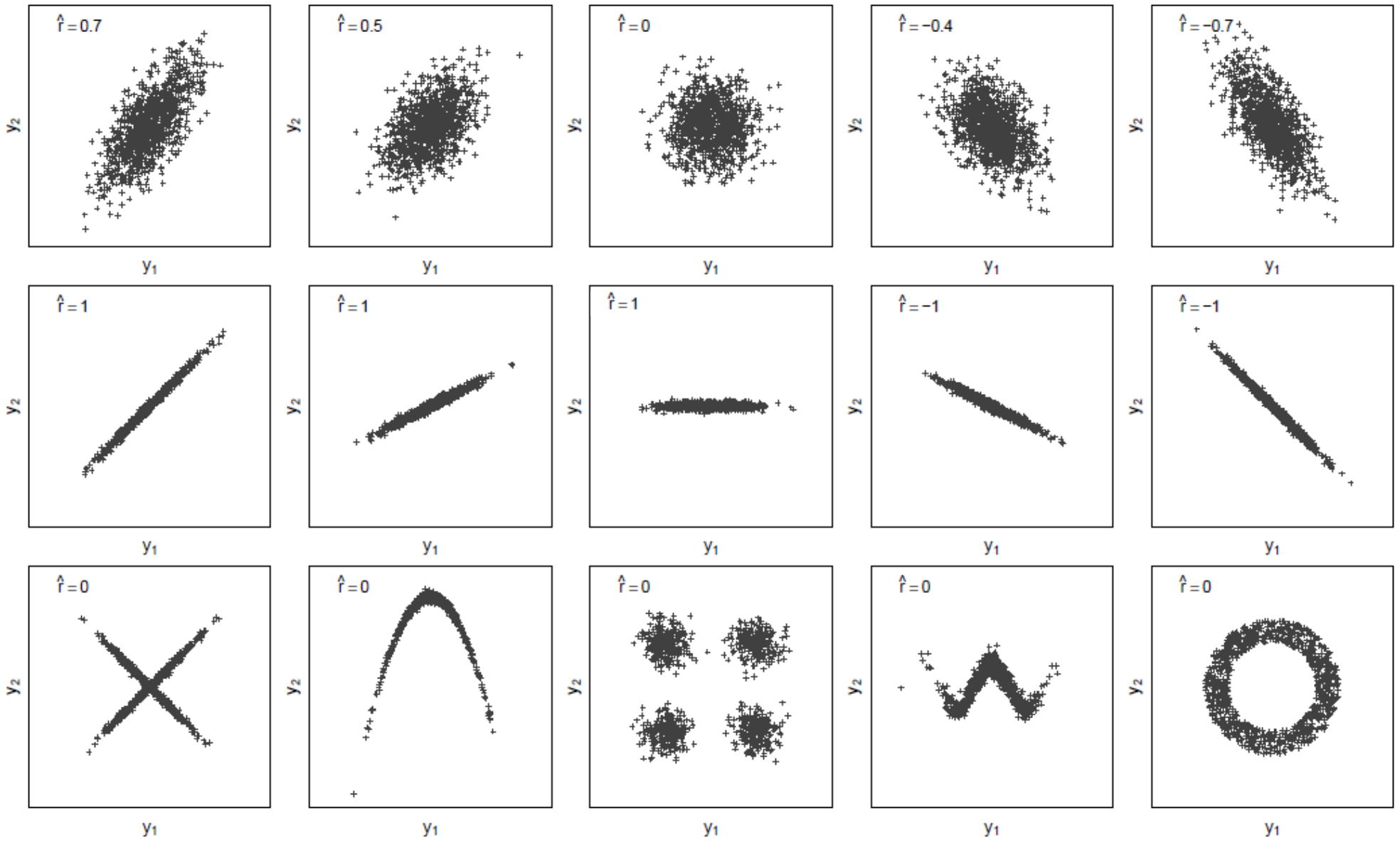
monotone non linéaire croissante

D
pas de corrélation, mais dépendance

non monotone



E
indépendance



Les deux premières lignes illustrent le fait que la corrélation ne mesure pas la pente de la relation linéaire entre les deux variable,

mais la dispersion autour de cette relation.

Il existe un **test** mais dont l'**utilité est très limitée** bien qu'il soit beaucoup appliqué. Il permet en effet uniquement de tester Ho : r=0 (i.e. corrélation nulle).

Or Ho peut être rejetée même pour une faible valeur de corrélation (par ex. r= 0,1).

On lit alors souvent dans la littérature que la « corrélation est significative » ce qui est généralement mal traduit ensuite comme « corrélation forte ».

C'est la valeur elle-même de la corrélation qui est importante à considérer.

La « **significativité** » (i.e. rejet de Ho) d'une correlation dépend du nombre d'observations.

Exemple :

- Un r = 0.6 établi sur un échantillon de 10 personnes n'est pas « significatif » au seuil de 5%, alors qu'une la corrélation est élevée.

- Un r = 0.2 établi sur un échantillon de 200 personnes est « significatif » au seuil de 5%, alors que la corrélation est faible.

→ Importance de considérer l'amplitude de la valeur de r !!

Le test permet uniquement d'évaluer si r est significativement différent de 0 !

Pour déterminer si une correlation est significativement différente de 0, il faut procéder à un **test d'hypothèse** :

Si n<100

(1) H0 : il n'y a pas de relation entre les deux variables X et Y, r = 0

(2) On fixe un risque d'erreur pour le rejet de H0 (ex. alpha=5%)

(3) On calcule la valeur absolue du coefficient de corrélation *r*(X,Y)

(4) On identifie la valeur théorique *r*(alpha, n-2 ddl) sur table Pearson

(5) On teste H0 vraie si r(alpha, d.d.l.) > abs(r(X,Y))

(6) On accepte ou rejette H0

Si n>100

r de Pearson peut être transformé en une variable F avec $\nu_1 = 1$ et $\nu_2 = n-1$ degrés de liberté.

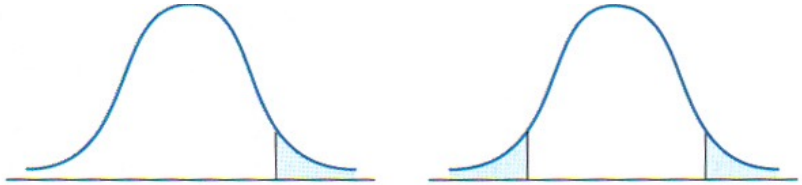
$$F = \frac{r_{xy}^2 / 1}{\left(1 - r_{xy}^2\right) / (n-2)} = \frac{r_{xy}^2 (n-2)}{1 - r_{xy}^2}$$

Puisque $\nu_1 = 1$, il est possible de transformer la statistique F en une de t avec $\nu = n-2$ degrés de liberté.

$$t = \sqrt{F} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$$

On rejette H_0 si $t_c \geq t_{(\alpha/2,n-2)}$ ou si $t_c \leq -t_{(\alpha/2,n-2)}$

Table de t de Student



α

Degrés de liberté

Valeur t

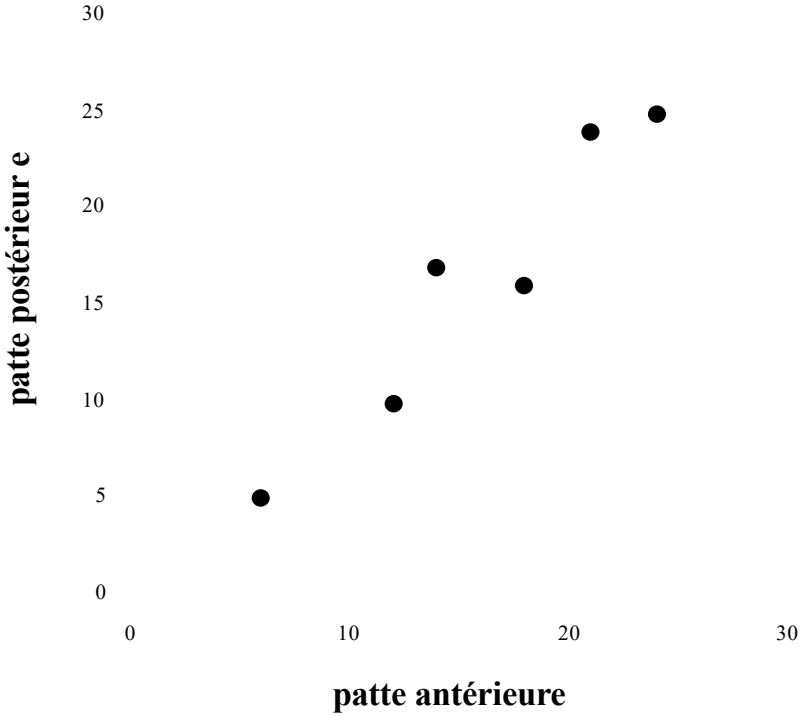
	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS COMBINED					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

Exemple d'application de la corrélation de Pearson

Existe-t-il une corrélation entre la longueur des pattes antérieures et postérieures chez les sciuridés ?

Espèce	antérieur	postérieur
1	6	5
2	12	10
3	14	17
4	18	16
5	21	24
6	24	25

-On observe une relation de type linéaire positive monotone croissante



1/ Calcul des écarts aux moyennes

	x	y	x-$\bar{x}$	y-$\bar{y}$
	6	5	-9.8	-11.2
	12	10	-3.8	-6.2
	14	17	-1.8	0.8
	18	16	2.2	-0.2
	21	24	5.2	7.8
	24	25	8.2	8.8
moyenne	15.8	16.2		

2/ calcul de r

$$r_{X,Y} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

→ $(-9.8)(-11.2)+(-3.8)(-6.2)+\ldots = 244.16$

→ $\sqrt{[(-9.8)^2+(-3.8)^2+\ldots]\,[(-11.2)^2+(-6.2)^2+\ldots]} = 139.50$

→ $r=0.912$

3/ On détermine la valeur r_{seuil} pour *n*-2 ddl et α=0,05

	Level of Significance *p* (two-tailed)			
df	0.10	0.05	0.02	0.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834

4/ On prend la décision statistique

$|r_{obs}| > r_{seuil}$ → H_0 rejetée

Il y existe une relation linéaire positive monotone croissante entre les 2 variables, elles sont positivement et fortement corrélées (r=0.91). r est significativement différent de 0.

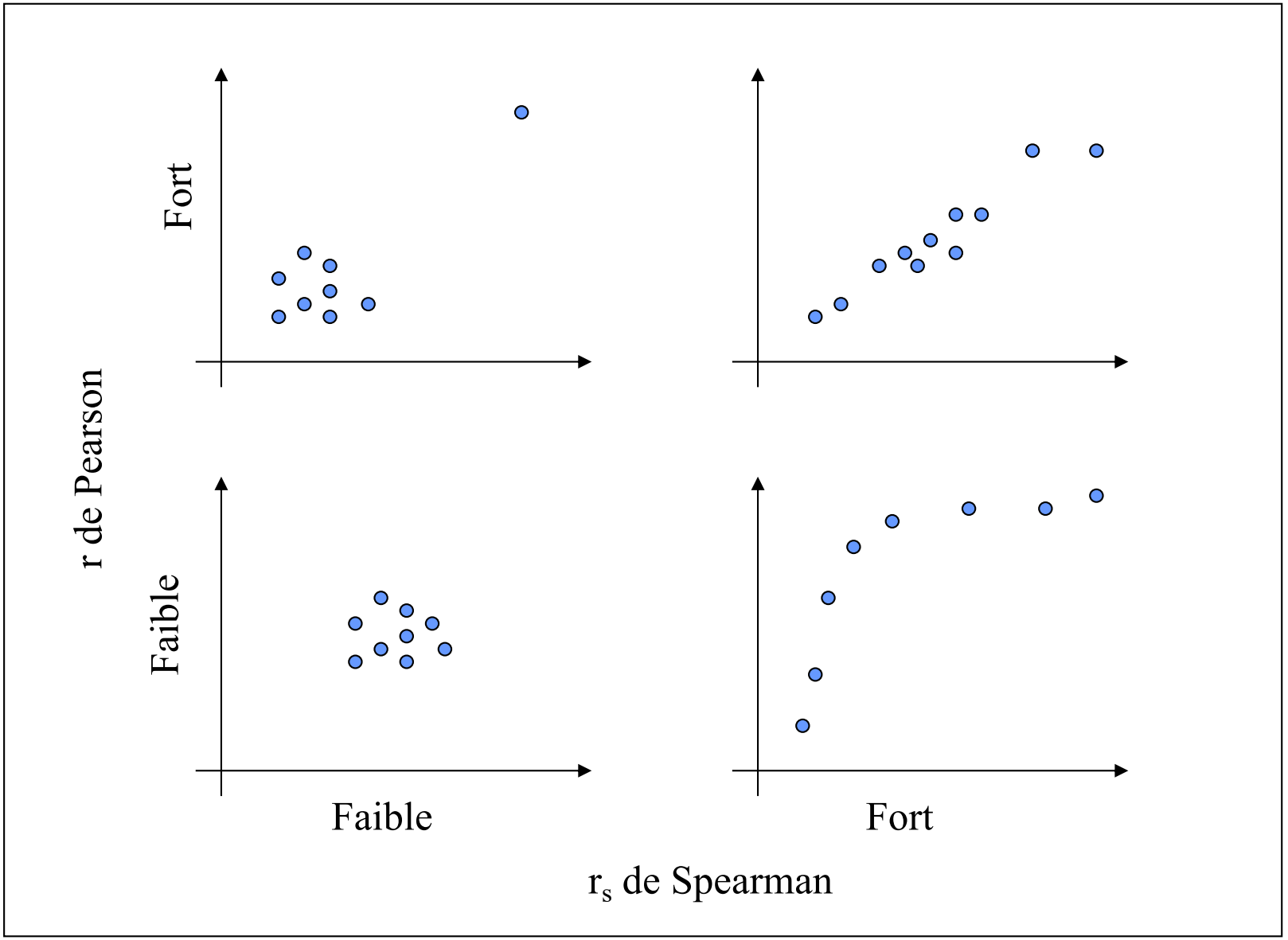
Coefficient de corrélation de Pearson, et test pour l'absence de correlation linéaire entre les variables

cor() et cor.test()

↳ Lien statistique significatif ≠ Lien de cause à effet

Coefficient et test valable uniquement pour des relations linéaires monotones, sans points extrêmes.

b) Corrélation non-paramétrique – r_s de Spearman



r_s :coefficient de corrélation de Spearman basé sur le rang des valeurs

$$r_s = 1 - \frac{6 \times \sum_{i=1}^{n} Di^2}{n(n^2 - 1)}$$

D représente, pour chaque observation, les différences de rang obtenues sur les deux variables.

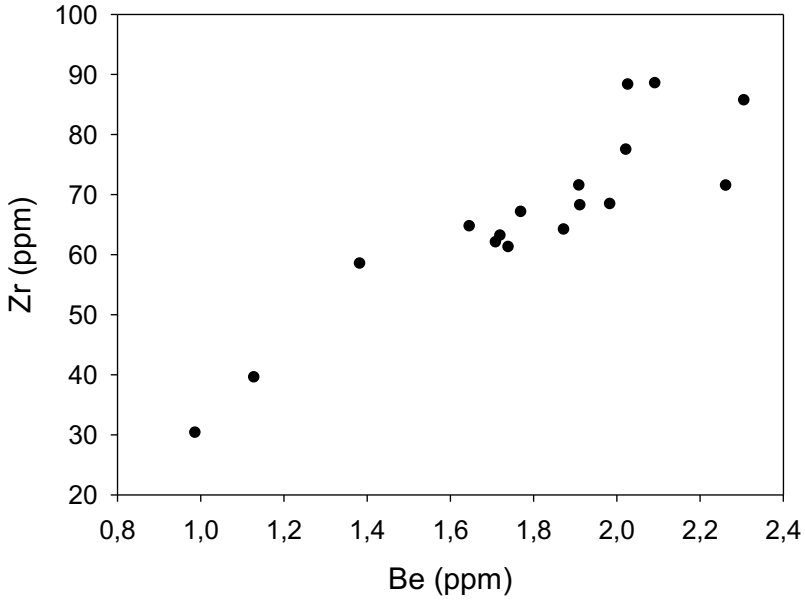
Exemple : relation entre concentrations de 2 métaux

Ech.	Be	Zr	Rang Be	Rang Zr	D²
1	1,71	62,04	5	5	0
2	1,91	71,50	10	13	9
3	1,98	68,40	12	11	1
4	1,74	61,25	7	4	9
5	1,87	64,16	9	7	4
6	1,38	58,49	3	3	0
7	0,99	30,33	1	1	0
8	1,13	39,55	2	2	0
9	1,65	64,71	4	8	16
10	2,26	71,47	16	12	16
11	1,72	63,14	6	6	0
12	1,77	67,09	8	9	1
13	2,31	85,68	17	15	4
14	2,09	88,52	15	17	4
15	2,03	88,30	14	16	4
16	2,02	77,45	13	14	1
17	1,91	68,20	11	10	1
			Somme D²		70

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

$$r_s = 1 - \frac{6 \times 70}{17 \times (17^2-1)}$$

$$r_s = 0.914$$



Cette valeur est-elle significativement différente de 0 ?

$H_0 : \rho_s = 0$ (absence de corrélation)

$H_1 : \rho_s \neq 0$

Deux cas possibles

Si n<100, table qui donne en fonction de n et α, la valeur r_{seuil} telle que sous H_0 on ait $P(|r_s| > r_{seuil}) = \alpha$

On rejette donc H_0 si $|r_s| > r_{seuil}$

Ici, n=17, r_s =0,914>0,485, donc H_0 est rejeté, la corrélation entre Zr et Be est donc significative différente de 0.

α / n	0.50	0.20	0.10	0.05
4	0.600	1.000	1.000	
5	0.500	0.800	0.900	1.000
6	0.371	0.657	0.829	0.886
7	0.321	0.571	0.714	0.786
8	0.310	0.524	0.643	0.738
9	0.267	0.483	0.600	0.700
10	0.248	0.455	0.564	0.648
11	0.236	0.427	0.536	0.618
12	0.224	0.406	0.503	0.587
13	0.209	0.385	0.484	0.560
14	0.200	0.367	0.464	0.538
15	0.189	0.354	0.443	0.521
16	0.182	0.341	0.429	0.503
17	0.176	0.328	0.414	0.485
18	0.170	0.317	0.401	0.472
19	0.165	0.309	0.391	0.460
20	0.161	0.299	0.380	0.447

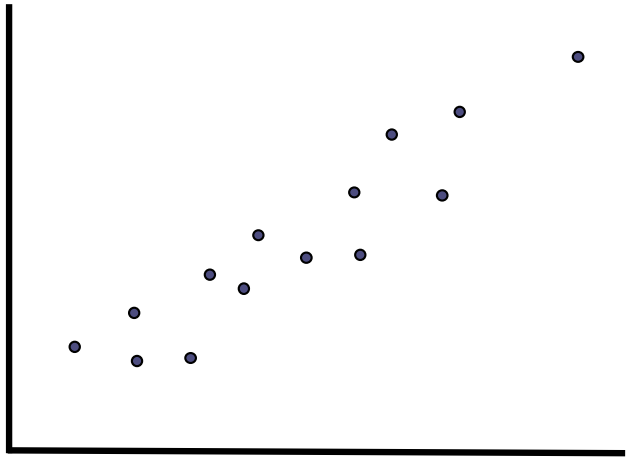
Si n>100

Calcul statistique t de Student (transformation de *r* en *F,* puis *t*) :

$$t_c = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

On rejette H_0 si $t_c \geq t_{(\alpha/2,n-2)}$ ou si $t_c \leq -t_{(\alpha/2,n-2)}$

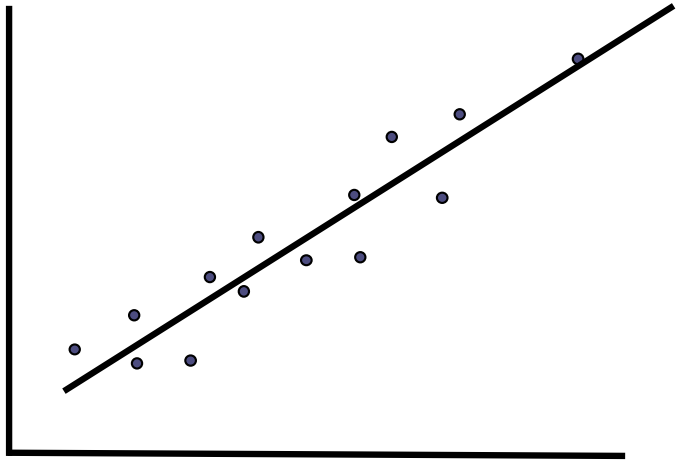
cor.test(Y1,Y2,method="spearman")



Corrélation entre 2 variables

(pente non calculée)

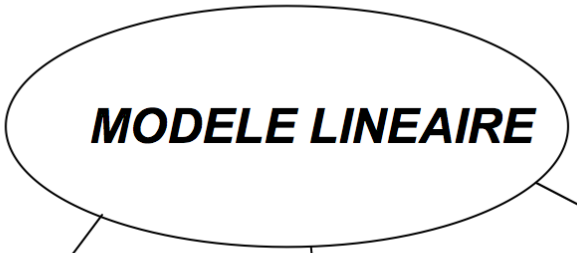
→ **Corrélation linéaire de Pearson**



Relation linéaire entre 2 variables et prédictions valeurs

(pente calculée)

→ **Régression linéaire**



ANOVA à 1 ou plusieurs facteurs

Variables dépendante : quantitative

Variable(s) indépendante(s) : qualitatives = des facteurs à plusieurs modalités

Plusieurs types d'ANOVA!

Régression simple ou multiple

Variable dépendante : quantitative

Variables indépendantes : quantitatives

ANCOVA

Variable dépendante : quantitative

Variables indépendantes : une combinaison de variables quantitatives et qualitatives

Conditions de validité de ces modèles linéaires sont identiques :
Normalité des résidus, homoscedasticité des résidus et indépendance des résidus