Master 2 PCPP

Cours Statistiques

STATISTIQUES DESCRIPTIVES 2
Distribution normale 3
STATISTIQUES INFÉRENTIELLES 5
TEST D'HYPOTHÈSE SUR UNE MOYENNE 6
Distribution de t 7
TESTS D'HYPOTHÈSE SUR LA DIFFÉRENCE ENTRE 2 MOYENNES 8
INTERVALLES DE CONFIANCE 10
PUISSANCE DE TEST 11
ANOVA À UN FACTEUR (*one-way*) 12
Comparaisons post-hoc 13
ANOVA À DEUX FACTEURS (*two-way*) 15
CORRÉLATION 17
Régression et prédiction 18
Interprétation de r 19
Coefficient de détermination 19
Inférences sur la significativité du r 20
CHI-2 (χ^2 ; *chi-square*) 21
Applications du χ^2 21
STATISTIQUES NON-PARAMÉTRIQUES 23
Coefficient de corrélation de Spearman (r_s) 24

Décembre 2020

L'étymologie ne nous apprend pas grand'chose : « status »… Utilisées dans le passé pour la collecte des impôts par les états, les Stats prennent une importance majeure dans la recherche moderne. Ex : en 1987, la FDA donne le feu vert pour la mise sur le marché de l'AZT en un temps record de 21 mois de recherche clinique (au lieu des ~9 ans habituels) étant donné la situation dramatique des victimes du SIDA. L'AZT avait des effets secondaires mais la preuve statistique d'une réduction du nombre de morts justifiait son utilisation.

On peut distinguer 2 sortes de Stats :

1) Stats descriptives : il s'agit d'organiser et résumer des observations. On ne fait pas de comparaisons et on s'intéresse en général à un seul groupe, échantillon ou population.
2) Stats inférentielles (ou inductives) : on peut ici viser 2 buts :
    a) Déduire les propriétés d'une population à partir de l'étude d'un échantillon. C'est par ex le principe des sondages. Il est important que l'échantillonnage soit fait au hasard (*random*). On met ici le doigt sur la notion de variabilité, principe inhérent à tout phénomène biologique.
    b) Comparer 2 ou plusieurs populations ou échantillons ; si une différence existe, on se demandera si cette différence est due à la variabilité (hasard), ou à un facteur différenciant les groupes étudiés.

Un troisième type de Stats à la charnière entre S descriptives et inférentielles a trait aux notions de corrélation et prédiction (voir chapitre concerné).

Dans toute démarche utilisant les Stats, il convient d'abord de poser une question « de recherche » (ex. AZT freine-t'elle la léthalité du SIDA ?), laquelle est différente de la question statistique où ce qui est traité, ce sont des données numériques. Les Stats font partie du plan (*design*) expérimental généré par la question de recherche. Ce plan fait en général intervenir 4 types de paramètres :

1) La variable indépendante : il s'agit du X, ex. le stimulus dans une étude stimulus-réponse ; exx. influence du stress dans un test de labyrinthe.
2) La variable dépendante : c'est Y, ce que l'on mesure, la réponse, le nombre de bons (ou mauvais) choix dans le labyrinthe.
3) Le ou les facteurs sujets d'étude : ex. effet d'un tranquillisant sur les relations entre stress et performance dans le labyrinthe.
4) Variables parasites : ex. coton autour du muscle en TP de LSV2 ; influence du cycle jour/nuit sur un dosage hormonal. Il faut faire en sorte que les variables parasites soient les mêmes pour tous les groupes.

Après un test, on tire une conclusion statistique d'ordre quantitatif (ex. il y a 5% de chances que tel résultat soit dû au hasard). Il ne s'agit pas d'une estimation qualitative : on ne peut pas dire par ex. que les groupes A et B <u>sont</u> différents. Après exécution du plan expérimental, lequel comprend plusieurs tests (parfois un grand nombre), on peut espérer atteindre à une conclusion « de recherche » d'ordre qualitatif.
Les Stats mentent-elles ? En dehors de la manipulation délibérée, la possibilité existe de faire des erreurs de « design », par ex en ne contrôlant pas certaines variables parasites ou en effectuant inconsciemment un échantillonnage non-aléatoire. D'autre part, la quasi-totalité des résultats publiés dans les journaux scientifiques sont des résultats « positifs » obtenus en général avec un seuil de significativité (*significance*) de 0,05. Cela signifie que si 20 équipes travaillent sur le même sujet de recherche, dont 19 ne trouvent pas de résultat positif, il existe 1/20 chances qu'un résultat « faux » soit publié… ! (ex des plannaires et des engrammes). Les erreurs d'échantillonnage sont les plus communes, particulièrement en rapport avec la taille. Une trop petite ou trop grande taille d'échantillon peut amener à des conclusions statistiques qui faussent la conclusion de recherche.

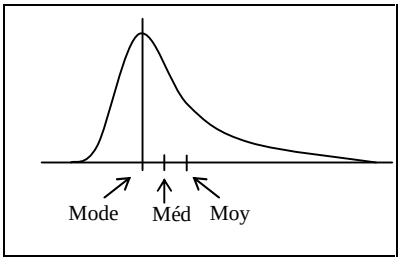
STATISTIQUES DESCRIPTIVES

Pour avoir un coup d'œil d'ensemble sur un grand nombre de données, on peut les représenter en distributions de fréquences, dont une forme commune est l'histogramme de fréquence. Dans ce dernier, le rapport de l'aire de chaque barre sur l'aire totale de l'histogramme donne la fréquence de l'intervalle par rapport au nombre total de cas dans la distribution. Un intervalle adéquat peut se calculer à partir de la formule de Sturge : $1+(3{,}3 \log_{10} n)$; ou de Yule : $2{,}5 \sqrt[4]{n}$. Différents types de fréquences peuvent s'exprimer :

1) Absolue
2) Relative : permet de comparer des groupes d'effectifs différents. Attention aux non-sens sur des n faibles (ex. le fait qu'un des 2 mécaniciens d'Aspremont soit alcoolique ne veut pas dire que 50% des mécaniciens d'Aspremont sont alcooliques)…
3) Cumulative absolue
4) Cumulative relative : permet de repérer les centiles (*percentiles*) d'une distribution. La courbe a une allure sigmoïde dont l'accélération centrale est due à la concentration des effectifs autour de la moyenne.

Trois paramètres suffisent à caractériser les distributions de fréquences :

1) Forme : Poisson (J inversé) ; asymétrique positive ou négative (*skewed*) ; rectangulaire ; bi- ou multimodale ; en cloche.
2) Tendance centrale
   a) Mode (NB : le mode ≈ la mode) : toujours utilisé avec les échelles nominales.
   b) Médiane : sépare l'effectif en 2 moitiés. Formule compliquée mais facile à repérer sur une distribution de fréquences cumulatives.
   c) Moyenne arithmétique : $\mu = \frac{\sum X}{N}$ pour la population ; $\overline{X} = \frac{\sum X}{n}$ pour l'échantillon. NB : i) $\sum(X-\overline{X}) = 0$. ii) La moyenne est sensible aux extrêmes de la distribution. iii) Est utilisée pour les tests statistiques si la distribution est normale car c'est le paramètre qui varie le moins d'un échantillon à l'autre. Dans une distribution asymétrique, la médiane est la meilleure représentation de la tendance centrale. iv) Dans une distribution symétrique, le mode, la médiane et la moyenne ont la même valeur.



   d) Moyenne géométrique de n valeurs : $n^{ème}$ racine de leur produit

   $MG = \sqrt[n]{\prod_{i=1}^{n} X_i}$; $\text{Log MG} = \frac{1}{n}(\log_{X1} + \log_{X2} + \ldots + \log_{Xn})$

3) Dispersion (variabilité)
Paramètre important pour les Stats inférentielles. Quantifiée par :

a) Etendue ou étalement (*range*) : max-min

b) Variance : comme $\sum(X-\overline{X}) = 0$, on prend le carré des déviations :

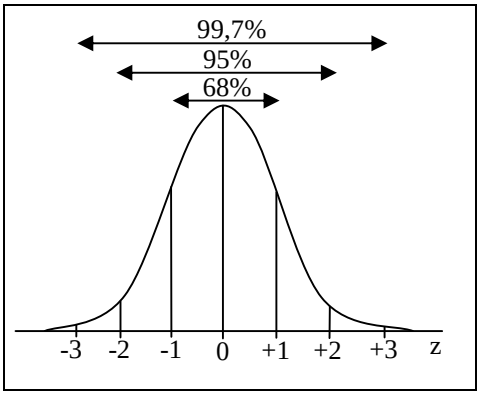
$\sigma^2 = \frac{\sum(X-\mu_X)^2}{N}$ (pop) ; $S^2 = \frac{\sum(X-\overline{X})^2}{n}$ (éch ; NB : avec n-1 au dénominateur, on a un estimateur non-biaisé de la variance de la population, s^2 → voir + loin). $\sum(X-\overline{X})^2$, la somme des carrés (SC) des déviations de X par rapport à la moyenne, est fréquemment utilisée en statistiques. Son calcul, potentiellement fastidieux, peut être simplifié par la formule suivante : $SC = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$. Ex :

	$X - \mu$	$(X - \mu)^2$
1	-3	9
5	+1	1
7	+3	9
3	-1	1
16/4 = 4	$\Sigma = 0$	$\Sigma = 20$; $\sigma^2 = 5$

c) Ecart-type (*standard deviation*) : $\sigma_X = \sqrt{\pm^2}$; $S_X = \sqrt{S^2}$

d) Ecart réduit (*z score*) : $z = \frac{X-\overline{X}}{S_X \,(ou\ s_X)}$; NB : $\mu_Z = 0$ et $\sigma_z = 1$.

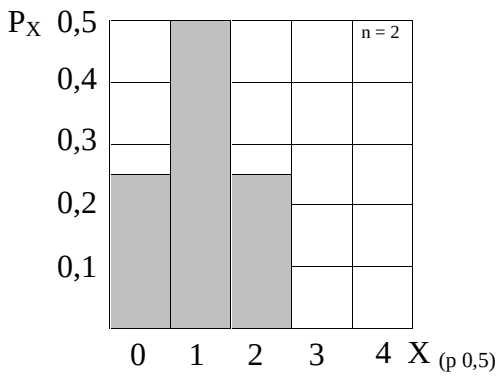
Distribution normale

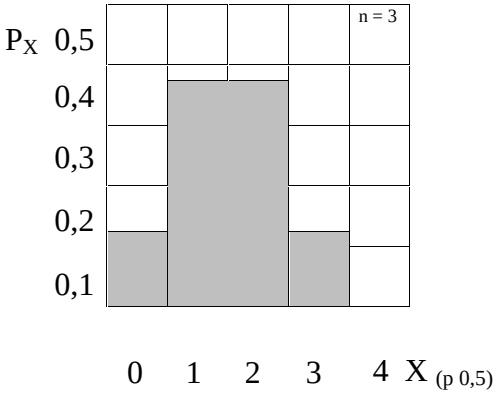


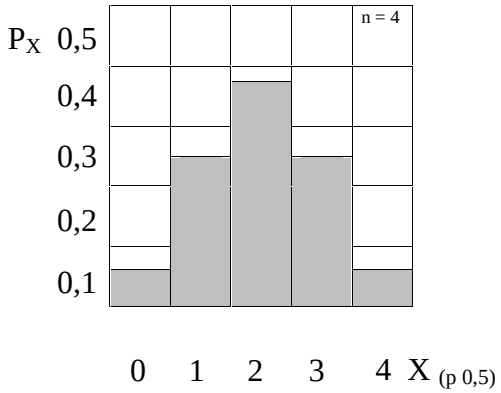
Propriétés : 95% des données sont comprises entre ± 1,96 et 99% entre ± 2,58 écarts-type. Ex : avec une moyenne et un écart-type de 100 ± 15, on sait que 95% des données sont comprises entre 70 et 130. On peut consulter une Table d'aire sous la courbe pour d'autres valeurs.
La courbe normale peut se décrire par un formalisme mathématique (sans grand intérêt ici) :

$Y = \frac{N}{\pm\sqrt{2®}}\ e^{-(X-\overline{X})^2/2\sigma^2}$

Représentation graphique







La forme vous rappelle quelque chose ?
→ La distribution normale est une distribution binômiale où p = 0,5. Beaucoup de tests statistiques sont basés sur l'hypothèse nulle, c-à-d qu'un résultat soit dû à la variabilité aléatoire. Comme on connaît la distribution de cette variabilité, on peut déterminer la proba (ex. 5%) qu'elle soit responsable de ce résultat, et retenir ou au contraire rejeter l'hypothèse nulle.

STATISTIQUES INFÉRENTIELLES

Buts : 1) Les caractéristiques de l'échantillon décrivent la population
2) Tester l'hypothèse nulle (H_0) qu'un résultat ou une différence entre groupes soient dûs au hasard. Exs : pile ou face sur 100 coups → si f s'éloigne trop de 0,5 la pièce est truquée ; résultats au bac du lycée Impérial comparés à la moyenne nationale ; drogue A comparée à drogue B ; etc…).

En fait les 2 buts sont liés dans les tests statistiques : dans l'exemple précédent des drogues A et B, on postule que les échantillons utilisés pour tester l'hypothèse sont représentatifs de leur population respective, ce qui permettra de prédire avec une certaine proba qu'il vaut mieux prescrire A, B, les 2 indifféremment, ou ni l'une ni l'autre. Un prérequis fondamental pour valider ce postulat est que les échantillons soient constitués de manière aléatoire.

Les tests d'hypothèse reposent essentiellement sur la mesure de la moyenne. Dans une population finie, si on extrait tous les échantillons possibles d'une taille donnée, on obtient une distribution de leurs moyennes appelée « distribution d'échantillonnage aléatoire de la moyenne » (*random sampling distribution of the mean*). La moyenne de ces moyennes, $\mu_{\bar{X}}$, est égale à la moyenne de la population : $\mu_{\bar{X}} = \mu_X$.

Soit une population [2,4,6,8] → 16 échantillons de 2 numéros tirés au hasard.
→ Proba de chaque échantillon = ¼ x ¼ = 1/16.
Fréquence (proba) de chaque moyenne :

8	1/16
7	2/16
6	3/16
5	4/16
4	3/16
3	2/16
2	1/16

val	X-μ	SC
2	-3	9
4	-1	1
6	+1	1
8	+3	9
5		20
$\sigma^2 = 20/4 = 5$		
$\sigma_X = \sqrt{5} = 2{,}24$		

Certaines populations sont finies (ex : loups des AM), mais dans la plupart des cas on a affaire à des populations infinies (ex : effets d'une drogue sur des rats). Pour une population infinie :

1) $\mu_{\bar{X}} = \mu_X$ (5 dans notre exemple)
2) $\pm_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$ → c'est l'écart-type de la distribution d'échantillonnage de $\bar{X}$, communément appelé erreur-type de la moyenne (*standard error of the mean* ; $2{,}24 / \sqrt{2} = 1{,}58$ dans notre exemple).
   NB : a) $\pm_{\bar{X}} < \sigma_X$
   b) $\pm_{\bar{X}}$ diminue quand σ_X diminue et quand n augmente
3) Théorème de la limite centrale : quand n augmente, la distribution des $\bar{X}$ tend vers une distribution normale quelque soit la distribution de la population d'origine (noter que la population [2,4,6,8] a une distribution rectangulaire, alors que la distribution d'échantillonnage des moyennes correspondantes (cf. tableau) se rapproche de la normale).

La distribution théorique des $\bar{X}$ permet de situer et comparer la moyenne d'un échantillon donné par rapport à cette distribution afin de retenir ou rejeter l'hypothèse que $\bar{X} = \mu_{\bar{X}}$.

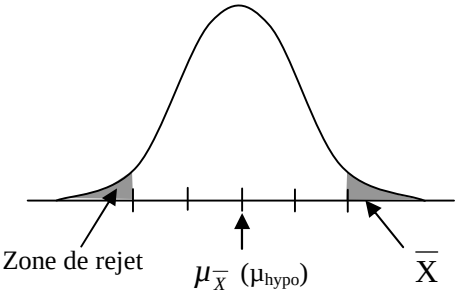
TEST D'HYPOTHÈSE SUR UNE MOYENNE

Ex: on veut déterminer si le niveau (noté sur 100) de sportifs niçois est différent de la moyenne nationale → H_0: $\overline{X} = \mu_X$
Niveau significatif $\alpha_{0,05} = \pm 0{,}025$

Moy obtenue ($\overline{X}$) $\mu_{\overline{X}}$ (μ_{hypo})

Ex : $z = \frac{\overline{X} - \mu_{hypo}}{\sigma_X/\sqrt{n}} = \frac{90-85}{20/\sqrt{100}} = \boxed{2{,}5}$

Ecart-type pop Taille éch.



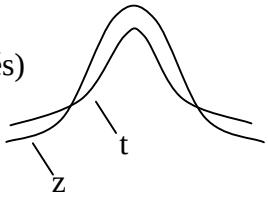
La moyenne obtenue est 2,5 erreurs-type au-dessus de la valeur attendue si H_0 était vraie.
→ H_0 est rejetée à $\alpha = 0{,}05$
→ … mais retenue (≠ acceptée) à $\alpha = 0{,}01$ ($z = 2{,}58$).
⇒ Importance de la taille de l'échantillon !
NB : il est le plus souvent impossible de connaître l'écart-type de la population entière, σ_X. Il faut donc l'estimer à partir de l'échantillon, comme on estime μ_X à partir de $\mu_{\overline{X}}$. Le problème est que la variance de l'échantillon est un estimateur biaisé de la variance de la population car S_X^2 est toujours inférieure à $\pm_{\overline{X}}^2$. La solution est de calculer s_X^2 (petit s) = $\frac{SC}{n-1}$, ce qui nous donne l'écart-type $s_X = \sqrt{s^2}$. On peut alors substituer σ_X par son estimateur non-biaisé s_X. Le nouveau dénominateur $s_{\overline{X}} = s_X/\sqrt{n}$ s'appelle l'« erreur-type estimée de la moyenne » (*estimated standard error of the mean*). Quand on substitue $s_{\overline{X}}$ à $\pm_{\overline{X}}$ dans la formule du z, on obtient le t de Student : $t = \frac{\overline{X} - \mu_{hypo}}{s_X/\sqrt{n}}$

A ce stade, un petit rappel ne sera probablement pas superflu…

σ^2 : variance de la population, $\frac{SC}{N}$
S^2 : variance de l'échantillon, $\frac{SC}{n}$
s^2 : estimateur non-biaisé de σ^2, $\frac{SC}{n\text{-}1}$
σ_X : écart-type de la population, $\sqrt{\sigma^2}$
S_X : écart-type de l'échantillon, $\sqrt{S^2}$
s_X : estimateur non-biaisé de σ_X, $\sqrt{s^2}$
$\pm_{\overline{X}}$: erreur-type de la moyenne, $\frac{\sigma_X}{\sqrt{n}}$
$s_{\overline{X}}$: erreur-type estimée de la moyenne, $\frac{s_X}{\sqrt{n}}$ ou $\frac{S_X}{\sqrt{n\text{-}1}}$ ou $\sqrt{\frac{s^2}{n}}$

Distribution de t

Même si $\overline{X}$ a une distribution normale, le fait de diviser par $s_{\overline{X}}$, qui n'est pas constant et varie d'un échantillon à l'autre, fait que t n'a pas une distribution normale.
→ Découverte de Gossett, qui écrivait sous le pseudo de Student (GB, ca. 1900).
Pour des échantillons dont n → ∞, t ~ z, sinon :



Similarités	Différences (des distributions)
- Moyenne = 0	- Leptokurtique (+ étroite au pic ; + large aux extrêmités)
- Symétrie	- $\sigma_t > \sigma_z$ (=1)
- Unimodicité	- Dépend du nombre de DL (t = z pour DL = ∞)

DL = degrés de liberté (*degrees of freedom*)
Ex : avec 5 DL, $t_{0,05}$ = 2,57 contre 1,96 pour z
→ Il faut « aller chercher » une différence plus loin avec t…
Pour consulter la Table de distribution de t, DL = n-1 car comme $\sum(X-\overline{X}) = 0$, le dernier $X-\overline{X}$ n'est pas libre de varier.

Ex. sur une pop de 3 éléments : ($s_X = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}}$; avec $\sum(X-\overline{X}) = 0$)

Si $X_1-\overline{X} = +3$ et $X_2-\overline{X} = -7$ → $X_3-\overline{X}$ = nécessairement +4.

Tests directionnels et non-directionnels (*one-tailed vs two-tailed*) : en choisissant une hypothèse alternative directionnelle (H_A: μ_X > ou < μ_{hypo}), toute la zone de rejet (ex : 5%) est reportée sur une des extrêmités (*tails*) de la distribution au lieu d'être répartie de chaque côté de la moyenne.
NB : il faut décider si on fait un test directionnel ou non-directionnel avant de recueillir les données. Sinon, si on fait par ex un test non-directionnel à $\alpha_{0,05}$ et qu'on passe ensuite à un test directionnel, on est passé en fait à $\alpha_{0,1}$. Idem pour pour le niveau de significativité : on doit déterminer α avant le recueil de données (mais voir NBB plus bas).

Erreur de type I : quand H_0 est rejetée alors qu'elle est vraie.
Erreur de type II : quand H_0 est retenue alors qu'elle est fausse.
NB : Erreur de type I = α (ex 0,05) = proba de rejeter H_0 quand elle est vraie.
NBB : au lieu de mentionner un seuil de significativité (ex p<0,05), on peut choisir de donner les valeurs exactes de $p_{erreur\ I}$ (ex $p = 3x10^{-5}$).

TESTS D'HYPOTHÈSE SUR LA DIFFÉRENCE ENTRE 2 MOYENNES

Même procédure / logique (H_0) que pour une moyenne unique, mais la distribution d'échantillonnage concerne maintenant toutes les différences de moyennes possibles entre 2 échantillons. Cette distribution a pour (1) moyenne « $\mu_{\bar{X}-\bar{Y}} = 0$ si H_0 est vraie.
Le théorème de limite centrale s'applique à cette distribution : elle est à peu près (2) normale même si les distributions de X et Y ne le sont pas.
A côté de la tendance centrale et de la forme, le 3ème paramètre qui caractérise une distribution est le degré de dispersion. La valeur de l'écart-type de la distribution des $\bar{X}-\bar{Y}$, ou erreur-type de la différence entre 2 moyennes, $\pm\sigma_{\bar{X}-\bar{Y}}$, va dépendre de la nature du test (éch. dépts contre indépts). Comme d'habitude, on va utiliser un estimateur non-biaisé, $s_{\bar{X}-\bar{Y}}$.
Le principe d'un test de différence de moyennes consiste à évaluer le rapport de cette différence à un écart-type estimé. On emploie le format général : « z » = $\frac{\Delta_\mu}{s}$, et on compare ce « z » à une valeur critique pour retenir ou rejeter H_0.

I. Cas d'échantillons indépendants (non-appariés)

1) Pour des variances inégales (hétéroscédasticité), $s_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}$, ce qui permet d'évaluer les variances (et le n) de chaque groupe indépendamment. La démarche est valable pour des échantillons de grande taille car le « z » suit alors une loi quasi normale. Pour de petits échantillons, le « z » n'obéit ni à une distribution normale ni à une distribution de t.
2) Pour remédier à ce problème, RA Fisher a introduit une modification qui consiste à mettre en commun la variance des 2 échantillons, ce qui génère un « z » qui suit une distribution de t. Cette modification est basée sur les prémisses (*assumption*) d'homogénéité des variances (homoscédasticité).
   Dans ce cas : $s_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_c^2}{n_X}+\frac{s_c^2}{n_Y}} = \sqrt{s_c^2(\frac{1}{n_X}+\frac{1}{n_Y})}$, où s_c^2 est la variance estimée commune ; $s_c^2 = \frac{\sum(X-\bar{X})^2+\sum(Y-\bar{Y})^2}{n_X+n_Y-2} = \frac{SC_X+SC_Y}{n_X+n_Y-2}$
3) Avec variances et n égaux, $s_{\bar{X}-\bar{Y}} = \sqrt{\frac{SC_X+SC_Y}{n(n-1)}} = \sqrt{\frac{2s_c^2}{n}}$

Selon le cas, on peut maintenant calculer notre t = $\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)_{hypo}}{s_{\bar{X}-\bar{Y}}}$

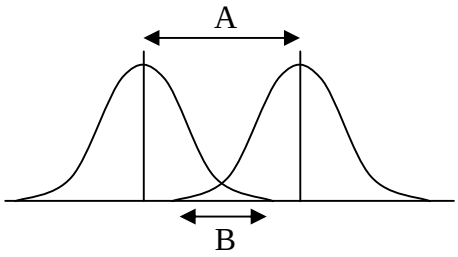
Dans le cas H_0: $\mu_X = \mu_Y$, t = $\frac{\bar{X}-\bar{Y}}{s_{\bar{X}-\bar{Y}}}$ avec DL = (n_X-1) + (n_Y-1).

Notas :
- Test directionnel ou non-directionnel : les mêmes principes que pour le test de moyenne unique s'appliquent ; le t calculé a la même valeur, mais la zone de rejet est soit répartie (non-directionnel) soit concentrée à un des 2 extrêmes (directionnel).
- Avec des σ^2 égales, le test est plus efficace (capable de détecter une différence si elle existe) quand $n_X = n_Y$ car alors $s_{\bar{X}-\bar{Y}}$ diminue.
- En principe l'homogénéité de variance est requise pour avoir une distribution de Student. Quand ce n'est pas le cas ($\sigma_X^2 \neq \sigma_Y^2$), on peut améliorer la situation en augmentant n (jusqu'à

~20/groupe), ou avec $n_X = n_Y$. Pour vérifier l'équivariance, on peut utiliser un test de F (voir plus loin) avec $F = \frac{s_1^2}{s_2^2}$, où $s_1^2 > s_2^2$, et DL = n_1-1 au numérateur et n_2-1 au dénominateur.
Si $F < F_{crit}$, on retient l'hypothèse nulle de l'égalité des variances.

- Quand les conditions sont trop défavorables, on peut utiliser des tests non-paramétriques (voir plus loin).

Principe général : A > B ?



II. Cas d'échantillons dépendants (appariés)

NB : l'appariement ne consiste pas seulement en mesures répétées sur le même groupe ; on peut aussi faire des paires sur la base d'un facteur commun (ex : QI dans un test sur le stress ou vice versa).

Ici $t = \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)_{hypo}}{\sqrt{s_{\overline{X}}^2+s_{\overline{Y}}^2-2r(s_{\overline{X}}s_{\overline{Y}})}}$, où r = coefficient de corrélation (voir plus loin) et $s_{\overline{X}}^2$ = carré de l'erreur-type.

Notas :

- Le principe d'équivariance n'est pas requis ici.
- DL = n <u>paires</u> – 1 car étant donné une valeur X, la valeur Y correspondante n'est pas complètement libre de varier.
- Dans un test apparié, on peut calculer $\overline{D}$, la moyenne des différences X-Y (contrairement au test non-apparié où on calcule la différence des moyennes), ainsi que leur écart-type estimé s_D. Sans avoir à calculer r, on peut obtenir le même résultat en testant H_0: $\mu_D = 0$, c'est-à-dire en comparant $\overline{D}$ à $\mu_{D(hypo)}$, ce qui revient à faire un test de moyenne unique :

$$\rightarrow t = \frac{\overline{D}-\mu_D}{s_D/\sqrt{n}}$$

INTERVALLES DE CONFIANCE

Dans de nombreux cas, l'estimation des IC s'avère plus utile et informative que les tests d'hypothèse. On ne s'intéressera ici qu'à l'estimation des IC de moyennes. Le but recherché consistera à calculer l'IC de la moyenne d'un échantillon comme l'intervalle de valeurs qui a 95 ou 99% de chances de contenir la moyenne de la population dont cet échantillon est extrait. Pour une distribution normale, si on connaissait σ_X cet intervalle pourrait être déterminé, grâce à l'écart réduit, comme $\overline{X} \pm z\,\sigma_{\overline{X}}$ (z = 1,96 ou 2,58 pour p = 0,05 ou 0,01 respectivement).

Ne connaissant pas la plupart du temps σ_X et $\sigma_{\overline{X}}$, on substitue $s_{\overline{X}}$ (= $s_X/\sqrt{n}$) comme estimateur de $\sigma_{\overline{X}}$. Ce faisant, t se substitue à z (voir plus haut), et on a $\overline{X} \pm t_p s_{\overline{X}}$, où p = 0,05 ou 0,01.

On cherche donc dans la Table le t correspondant avec DL = n-1 (NB : quand n↑, IC↓).

On obtient l'IC en ajoutant puis en soustrayant $t_p s_{\overline{X}}$ à $\overline{X}$. On est alors 95 ou 99% confiant que cet IC contient μ_X.

De même, on peut calculer l'IC d'une différence entre 2 groupes, $(\overline{X}-\overline{Y}) \pm t_p s_{\overline{X}-\overline{Y}}$, avec $s_{\overline{X}-\overline{Y}}$ calculé différemment selon qu'on a des échantillons dépendants ou indépendants.

On peut aussi exprimer l'IC en nombre d'écarts-type de la variable :

Pour une moyenne unique, la différence entre $\overline{X}$ et les limites de l'IC, $d_1 = \frac{t_p s_{\overline{X}}}{s_X} = \frac{t_p}{\sqrt{n}}$

Ex : $\overline{X} = 85$; $s_X = 15$; $n = 25$ → $s_{\overline{X}} = \frac{15}{\sqrt{25}} = 3$; $t_{0,05} \approx 2$ avec DL = 24.

→ 85 – (2x3) = 79 ; 85 + (2x3) = 91

$d_1 = 6/15 = 0,4$; $\frac{t_p}{\sqrt{n}} = 2/5 = 0,4$.

Pour une différence entre 2 moyennes : $d_2 = \frac{t_p s_{\overline{X}-\overline{Y}}}{s_{moy}}$, où s_{moy} est la moyenne de s_X et s_Y.

<u>Relations de l'IC avec H_0</u> :

Attention ! Il peut y avoir une différence significative entre 2 groupes dont les IC se chevauchent (*overlap*).

D'autre part, l'IC d'une différence entre 2 groupes peut éventuellement comporter une valeur négative et une valeur positive dont les positions relatives par rapport au zéro déterminent l'interprétation de l'IC (ex : 5±10). Néanmoins, la présence du zéro dans l'IC_{dif} permet de retenir H_0, alors qu'inversement si l'IC_{dif} ne contient pas zéro, on peut rejeter H_0.

Ex : $\overline{X}-\overline{Y} = 12$; $s_{\overline{X}-\overline{Y}} = 5$; n = 25x2 → $t = 12/5 = 2,4 > t_{crit} = \pm 2$ avec $\alpha = 0,05$ et DL = 48.

$IC_{95} = 5x2 = 10$ → $2 \leq \overline{X}-\overline{Y} \leq 22$, où zéro n'apparaît pas…

En conclusion, l'IC peut permettre de détecter une différence en plus de l'information spécifique qu'il apporte. La méthode apparaît donc supérieure dans bien des cas, en particulier quand il s'agit d'estimer la variabilité d'un paramètre. En général, on choisira la méthode H_0 quand il y a une décision à prendre…

PUISSANCE DE TEST
(+ détermination de la taille de l'échantillon)

Comme un trop petit échantillon peut faire rater une différence importante, un trop grand échantillon peut révéler une différence sans importance bien qu'elle soit significative.
Ex : QI / taille des enfants corrélés (calculé a posteriori, r = 0,03 !) à $p<0,001$ avec un échantillon de 14000 (NY Times, 1986).

Rappel : erreur de type I = α = proba de rejeter H_0 quand elle est vraie
- Erreur de type II = β = proba de retenir H_0 quand elle est fausse

→ 1 - β = proba de rejeter H_0 quand elle est fausse (= proba de détecter une différence) = puissance du test.
NB : quand β diminue la puissance du test augmente…

Facteurs qui affectent la puissance du test :

1) Distance $\overline{X}$ → μ_{hypo} (plus la distance est grande, plus il y a de chances de rejeter H_0)
2) Taille de l'échantillon : $s_{\overline{X}} = s_X/\sqrt{n}$ → plus n est grand, plus $s_{\overline{X}}$ est petit, moins il y a de chevauchement entre les distributions des moyennes $\overline{X}$ et μ_{hypo}.
3) Ecart-type des distributions, s_X → éviter les variables parasites et essayer d'apparier…
4) Niveau de significativité : plus α augmente plus la puissance de test augmente (au détriment de la protection d'une erreur de type I).
5) Un test unidirectionnel est plus puissant qu'un test bidirectionnel.

Résolution (*effect size*)
Il s'agit ici d'une question de recherche (≠ statistique) : quelle est l'ordre de grandeur d'une différence (*discrepancy* - entre la valeur d'hypothèse et la valeur vraie) qu'il nous semble important de considérer ? Une manière pratique de répondre à cette question est d'exprimer cette différence (d) en termes d'écarts-type (estimés le plus souvent).

Pour un test d'hypothèse sur une moyenne unique, $d = \frac{\overline{X} - \mu_{hypo}}{\sigma \text{ (ou } s)}$

Pour un test d'hypothèse sur une différence de moyennes, $d = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{hypo}}{\sigma \text{ (ou } s)}$

On peut par ex. considérer respectivement comme modérée et conséquente une d de 0,5 et 0,8 écarts-type.
NB : on ne peut en principe calculer la puissance du test que si μ est connue, ce qui est rarement le cas → choisir une puissance d'au moins 0,8 (80% de chances de rejeter H_0 quand elle est fausse).
→ On peut maintenant déterminer la taille de l'échantillon en fonction de la :
- Puissance du test = proba de ne pas rater une…
- Différence de grandeur choisie (résolution).

Par ex., avec d = 0,5 et 1- β = 0,8, la consultation de courbes de puissance (*power curves*) pour un test de t non-apparié et non-directionnel à α = 0,05 donne un effectif (n) de ~60.

		Réalité	
		H_0 fausse	H_0 vraie
Décision	Rétention	Type II	OK
	Rejet	OK	Type I
		Domaine β	Domaine α

ANOVA À UN FACTEUR (*one-way*)

Pourquoi ne pas faire n tests de t ?
- Lourd ! (ex : avec 7 groupes → 21 tests)
- Augmentation dramatique du risque d'erreur de type I : avec $\alpha = 0,05$ et des groupes indépendants, la proba d'erreur = $1\text{-}0,95^n = 1\text{-}0,95^{21} = 0,66$ avec 21 tests !
- Avec n groupes, on ne sait pas à quoi au juste on compare un groupe donné…

C'est Fisher, père de la statistique moderne, qui développe l'ANOVA.
H_0: $\mu_A = \mu_B = \mu_C = \ldots = \mu_k$
NB : Pas de test directionnel avec ANOVA puisque H_0 peut être fausse d'un grand nombre de façons…
NBB : H_A implique que les différents « traitements » peuvent représenter différentes « populations ».

I. Groupes indépendants

1) Variation intra-groupe = aléatoire (inhérente) → devrait être la même pour chaque groupe.
2) Variation inter-groupes → due aux différences parmi les moyennes ET à la variation inhérente.

En supposant pour simplifier que $n_A = n_B = n_C = \ldots = n_k$: $\frac{\sum s^2}{k} \xrightarrow{\text{estime}} \sigma^2$ (intra)

→ La variance moyenne des k groupes est une bonne estimation de la variance de la population sous-jacente qui ne dépend pas des $\overline{X}$ ni de H_0.
Par contre, si les groupes viennent de la même population (H_0 vraie), comme $s^2_{\overline{X}} = s^2/n$, on peut dire : $n\, s^2_{\overline{X}} \xrightarrow{\text{estime}} \sigma^2$ (inter), où $s^2_{\overline{X}}$ est la variance des moyennes de chaque groupe.
On a donc 2 estimations de σ^2. Si H_0 est vraie, les 2 estimations devraient être semblables.
Si H_0 est fausse, $\sigma^2_{inter} > \sigma^2_{intra}$. Ceci constitue le principe de base de l'ANOVA.

Procédure :
- Calculer la moyenne générale $\overline{\overline{X}}$, puis partitionner en écarts par rapport à $\overline{X}$ et $\overline{\overline{X}}$:

$$X - \overline{\overline{X}} = (X - \overline{X}) + (\overline{X} - \overline{\overline{X}})$$

- Calculer SC totale :

$$SC_t = \sum(X - \overline{\overline{X}})^2$$

→ $\sum(X - \overline{\overline{X}})^2 = \underbrace{\sum(X - \overline{X})^2}_{SC_{intra}} + \underbrace{\sum^k n_i (\overline{X}_i - \overline{\overline{X}})^2}_{SC_{inter}}$ où n_i = effectif de chaque (k) groupe et $\overline{X}_i$ = moyenne de chaque groupe.
Concrètement, la partition de SC_t en SC_{intra} (variation indépendante du « traitement ») + SC_{inter} (variation indépendante + dépendante du traitement) se fait :

$SC_{intra} = \sum(X - \overline{X})^2 = \sum(X_A - \overline{X}_A)^2 + \sum(X_B - \overline{X}_B)^2 + \ldots$ k

$SC_{inter} = \sum^k n_i (\overline{X}_i - \overline{\overline{X}})^2 = n_A (\overline{X}_A - \overline{\overline{X}})^2 + n_B (\overline{X}_B - \overline{\overline{X}})^2 + \ldots$ k

Degrés de liberté :
Rappel : $s^2 = SC/DL$ (= n-1 pour 1 échantillon)
DL pour $SC_t = n_t-1$

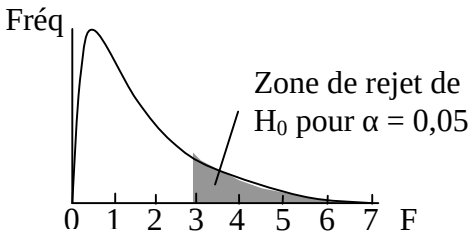
$SC_{intra} = (n_A - 1) + (n_B - 1) + ... = n_A + n_B + ... + n_k - k = n_t - k$

$SC_{inter} = k-1$

NB : $n_t - k + k-1 = n_t-1$ = DL pour SC_t

→ $s^2_{intra} = \frac{SC_{intra}}{n-k}$; $s^2_{inter} = \frac{SC_{inter}}{k-1}$, et $\boxed{F = \frac{s^2_{inter}}{s^2_{intra}}}$

→ Comparer ensuite à la valeur critique de F dans la Table avec n DL au numérateur et n DL au dénominateur.
Le F suit des distributions asymétriques qui varient selon le nombre de DL.
Ex. pour 4/20 et H_0 vraie :



NB : - F toujours >0 car il ne peut y avoir de variance négative
- Si F<1 → suspecter un problème…
- Pour calculer SC, utiliser plutôt la formule $\Sigma X^2 - \frac{(\Sigma X)^2}{n}$

Prérequis (*assumptions*) :
- Distribution normale (sinon OK quand n est grand)
- Homoscédasticité (sinon OK quand les groupes ont la même taille)
- Indépendance vraie pour ce type d'ANOVA
- Echantillonnage aléatoire

On a vu plus haut qu'on pouvait utiliser le F pour tester l'homoscédasticité de 2 échantillons. Pour >2 échantillons, il faut utiliser d'autres tests tels que Bartlett ou Levene (voir Scherrer).

Comparaisons post-hoc

A n'utiliser que si F est significatif…
En ordre croissant de « conservatisme » (donc décroissant de puissance) :
Duncan, Newman-Keuls, HSD de Tukey, Scheffé, etc.

1) Pour comparer toutes les paires possibles, on peut utiliser le HSD de Tukey : on compare la différence de 2 moyennes (2 groupes) à HSD (honestly significant difference). Si $\overline{X}_A - \overline{X}_B > HSD$, on rejette H_0.

   $HSD = q\sqrt{\frac{s^2_{intra}}{n}}$, où q est localisé dans une Table en fonction de k, DL et α.

   Si n varie d'un groupe à l'autre, on utilise la moyenne harmonique de n :

   $\tilde{n} = \frac{k}{(1/n_A) + (1/n_B) + ... + (1/n_k)}$

   NB : alternative à ANOVA : comparaisons plannifiées, non-développé ici...

2) Pour comparer chaque groupe à un témoin, on peut utiliser Dunnett :
H_0: $\overline{X}_A = \overline{X}_{témoin}$; $\overline{X}_B = \overline{X}_{témoin}$; etc…

$t = \frac{\overline{X}_A - \overline{X}_{témoin}}{\sqrt{s^2_{intra}(\frac{1}{n_A} + \frac{1}{n_{témoin}})}}$ est localisé dans une Table pour α, k (témoin inclus) et DL_{intra}.

II. Groupes dépendants (mesures répétées)

Uniquement avec des groupes de même taille !

SC_t et SC_{inter} ne changent pas (SC_{inter} est plus facile à calculer car n est le même).

SC_{intra} est partitionnée entre : - SC_{el} : variation entre éléments et
- SC_{res} : variation résiduelle (= aléatoire).

Ex :

	A	B	C	
1	1	2	3	6
2	2	3	4	9
3	3	4	5	12
	Variation éléments			

	A	B	C	
1	1	2	3	6
2	2	1	3	6
3	3	2	1	6
	Variation résiduelle			

$SC_{el} = \Sigma\, k\, (\overline{X}_{el} - \overline{\overline{X}})^2$, où $\overline{X}_{el}$ = la moyenne d'un élément dans k conditions (traitements)

$SC_{res} = SC_t - SC_{inter} - SC_{el}$ avec $DL_{el} = n-1$ et $DL_{res} = (DL_{el})(DL_{inter}) = (n-1)(k-1)$

$$\boxed{F = \frac{s^2_{inter}}{s^2_{res}}} = \frac{\text{variation aléatoire} + \text{effet traitement}}{\text{variation aléatoire}}$$

NB : ce test est plus puissant que pour les groupes indépendants car on enlève la variation due aux différences entre éléments : $s^2_{res} < s^2_{intra}$

On peut aussi utiliser les tests post-hoc pour mesures répétées, mais :

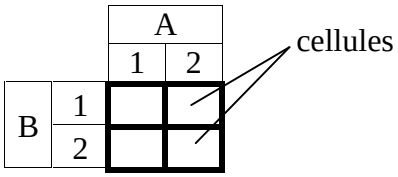
$HSD = q\sqrt{\frac{s^2_{res}}{n}}$ (au lieu de s^2_{intra}).

ANOVA À DEUX FACTEURS (*two-way*)

Ex : effets des engrais A_1 et A_2 (mesurés par le rendement) sur les variétés de blé B_1 et B_2.

I. Effets principaux

1) Effet variété → H_0: $\overline{X}_{B1} = \overline{X}_{B2}$
2) Effet engrais → H_0: $\overline{X}_{A1} = \overline{X}_{A2}$



→ Revient à faire des ANOVA à 1 facteur : $F_1 = \frac{s^2_A}{s^2_{res}}$; $F_2 = \frac{s^2_B}{s^2_{res}}$, où la variance résiduelle (inhérente) correspond ici à la variance intra-cellule (voir plus bas).

II. Interaction : les effets d'un facteur influencent-ils ceux de l'autre ? → $F_3 = \frac{s^2_{AB}}{s^2_{res}}$

Partition des variances :
Dans l'ANOVA avec mesures répétées on partitionnait SC_{intra} en SC_{el} et SC_{res}. Ici on partitionne SC_{inter} en $SC_{lignes} + SC_{colonnes} + SC_{lxc}$.
→ On a 5 SC :

- SC_t : identique à ANOVA à 1 facteur (variation de tous les X par rapport à $\overline{\overline{X}}$)
- $SC_{intra\text{-}cellules} = \Sigma (X - \overline{X}_{cell})^2$
- $SC_{lignes} = \Sigma n_{lignes} (\overline{X}_{lignes} - \overline{\overline{X}})^2$ → identique à SC_{inter} dans ANOVA-1 facteur où chaque ligne est considérée comme un ensemble.
- $SC_{col} = \Sigma n_{col} (\overline{X}_{col} - \overline{\overline{X}})^2$ → identique à SC_{inter} dans ANOVA-1 facteur où chaque colonne est considérée comme un ensemble.
- $SC_{lxc} = SC_t - (SC_{intra\text{-}cell} + SC_{lignes} + SC_{col})$

Attention : n_{lignes} et n_{col} représentent le nombre de X par lignes et par colonnes.

Degrés de liberté :
$DL_t = n_t - 1$; DL_{col} = nombre de colonnes-1 ; DL_{lignes} = nombre de lignes-1 ;
$DL_{intra\text{-}cell} = \Sigma (n_{intra\text{-}cell} - 1)$; DL_{lxc} = (nombre de lignes-1) (nombre de colonnes-1)
→ On peut calculer les variances estimées intra-cellule, lignes, colonnes et lxc par le rapport SC/DL correspondant.

- $s^2_{intra-cell}$ estime la variation inhérente
- s^2_{lignes} estime la variation inhérente + effet principal de la variété de blé
- s^2_{col} estime la variation inhérente + effet principal de la variété d'engrais
- s^2_{lxc} estime la variation inhérente + effet de l'interaction

Pour déterminer les 3 effets (A, B et AxB) on calcule 3 F :

$F_A = \frac{s^2_{col}}{s^2_{intra\text{-}cell}}$; $F_B = \frac{s^2_{ligne}}{s^2_{intra\text{-}cell}}$; $F_{AB} = \frac{s^2_{lxc}}{s^2_{intra\text{-}cell}}$.

On les compare ensuite au $F(H_0)$ de la Table avec les DL appropriés. Si $F > F_{Table}$ → H_0 rejetée.

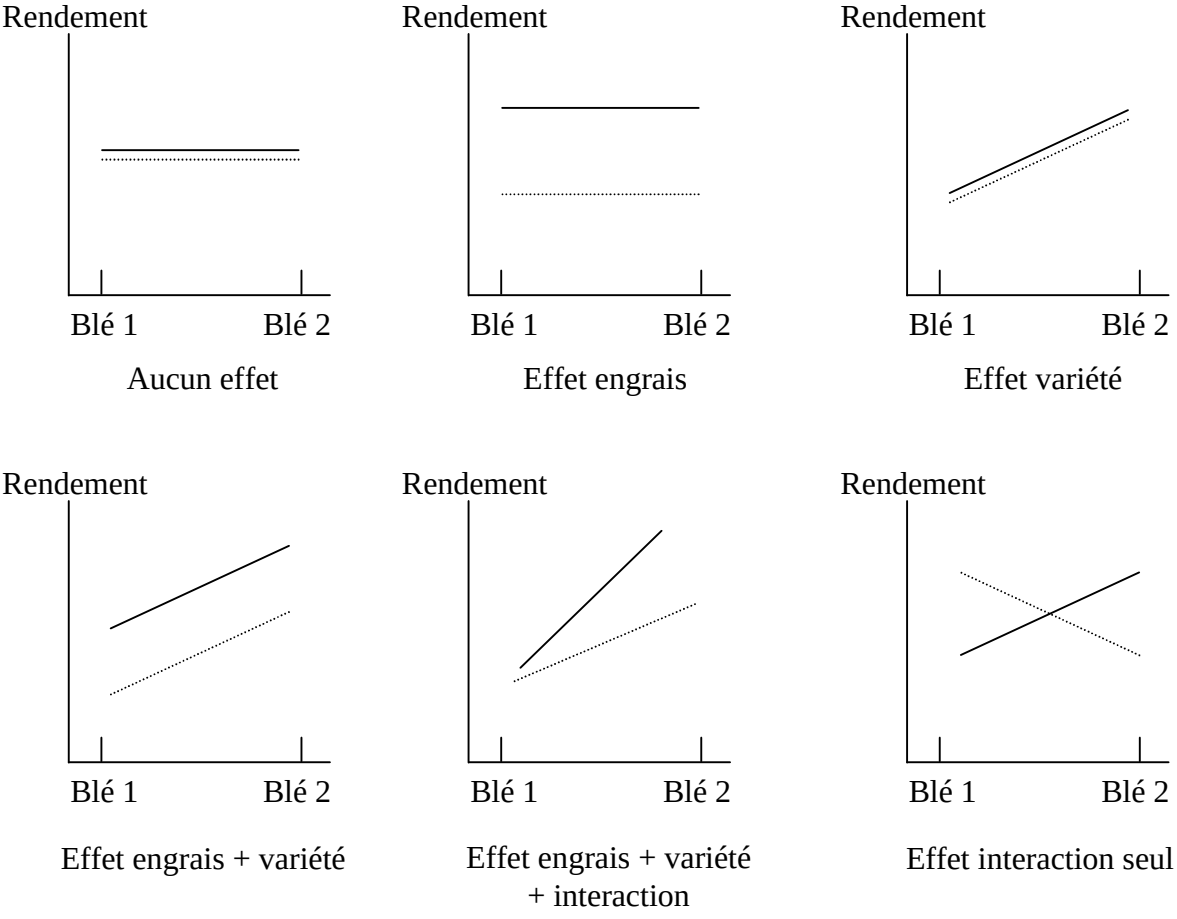
On peut enfin faire des comparaisons post-hoc si le nombre de conditions par facteur est >2.

Ex : HSD = $q\sqrt{\frac{s^2_{intra\text{-}cell}}{n}}$ où q est localisé dans la Table en fonction de α, du nombre de conditions par facteur et de $DL_{intra\text{-}cell}$, et où n = nombre de X par colonnes ou par lignes, selon qu'on cherche les différences parmi les conditions des facteurs A ou B.
→ Une différence de moyenne entre 2 conditions d'un facteur donné est significative si elle est >HSD.

Notas :
- Pour interpréter l'effet interaction, il faut en principe corriger la moyenne de chaque cellule en soustrayant les effets principaux :

moy. cell. corrigée = moy. cell. - $((\bar{L} - \bar{\bar{X}}) + (\bar{C} - \bar{\bar{X}}) + \bar{\bar{X}})$, où $\bar{L}$ et $\bar{C}$ sont les moyennes de chaque ligne et colonne qui constituent la cellule.
- Les prérequis sont les mêmes que pour ANOVA à 1 facteur, sauf qu'ici un nombre égal de X par cellule est plus important.
- On peut aussi faire des ANOVA à 2 (ou +) facteurs avec mesures répétées. Ex des contrôleurs aériens : effets de 2, 12 ou 24 h de veille sur un test de vigilance répété lors de 4 sessions sur n sujets. Ce type d'analyse est assez complexe → voir ouvrages spécialisés…

Exemples de résultats avec ANOVA-2F :



CORRÉLATION

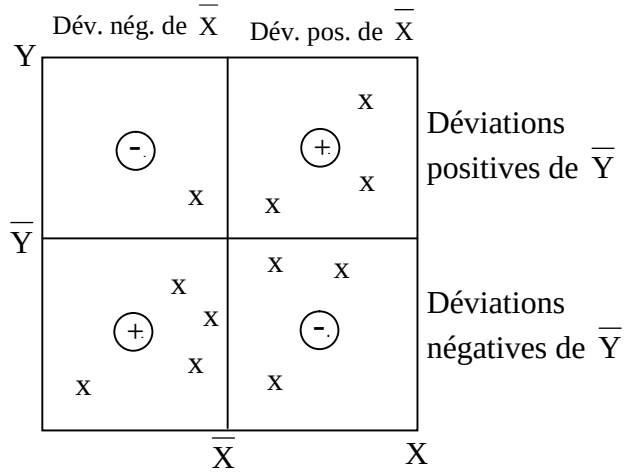
Notion liée à celle de prédiction. Exs : taille parents / enfants (Galton) ; nombre de cigarettes fumées / incidence du cancer pulmonaire ; ex. de corrélation négative : nombre d'heures passées par des enfants devant la télé / résultats aux tests de lecture…
On part d'une distribution bidimensionnelle (*bivariate*) ; on fait d'abord un graphe (ou diagramme) de dispersion (*scatterplot*). Ceci permet de repérer a priori une relation <u>linéaire</u>, la seule accessible au traitement mathématique de Pearson (~1900).
NB : il existe aussi des relations non-linéaires (ex : courbes dose-réponse ; courbes en cloche, etc.). Dans ce cas, il faut mathématiquement linéariser les données (ex. log X si Y = a log X + b)
La corrélation peut être positive ou négative, la pente de la relation étant toujours ≠ 0.

Coefficient de corrélation (de Pearson) : $-1 \leq r \leq +1$

$r = \frac{\sum(z_X z_Y)}{n} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{nS_XS_Y}$ où n = nombre de <u>paires</u> de valeurs.

Finalement, $r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{(SC_X)(SC_Y)}}$

NB : les $(X-\overline{X})$ et $(Y-\overline{Y})$ peuvent être positifs ou négatifs, de sorte que la somme des produits + et - détermine le signe de r (corrélation positive ou négative), le dénominateur étant toujours positif.



Notas :
- Prendre un échantillon assez grand sur le plan taille absolue (N) et sur le plan « gamme de talents ».
- Corrélation ≠ causalité ; ex : X ↔ Y ? ou Z (↙ X, ↘ Y) ?

Ex : force physique / intelligence chez les enfants de 8 à 12 ans ; nombre de nids de cigognes / nombre de naissances dans les villes d'Europe…
- Il existe d'autres types de corrélation, ex. Spearman (voir dernier cours sur tests non-paramétriques).

Régression et prédiction

Le terme de régression vient du fait que pour des valeurs extrêmes de X, les Y ont tendance à « régresser » vers la moyenne. Par ex., les enfants de parents de taille ou de QI très élevés ont tendance à être plus grands et intelligents que la moyenne, mais pas autant que leurs parents. Les étudiants très doués en maths ont tendance à être forts en stats, mais ne sont pas forcément les meilleurs… Le degré de régression vers la moyenne dépend du r : nul si r=1, total si r=0.
On trouve le meilleur ajustement (*best fit*) de la droite de régression par la méthode des moindres carrés (*least squares*) : si $d_Y = Y'-Y$ (où Y' = valeur prédite de Y), $\sum d_Y^2$ doit être le plus petit possible.
Dans le cas où on a plusieurs valeurs de Y pour chaque X, on prédit des valeurs Y' qui seront différentes des $\overline{Y}_X$ (moyennes des Y pour un X donné), et qui en sont des estimations. Dans ce cas, $d_Y = Y'- \overline{Y}_X$
NB : on peut aussi faire une régression de « X en Y », c-à-d prédire X en minimisant $\sum d_X^2$.
Sauf dans le cas où r = ±1, les droites de régression de Y en X et de X en Y seront différentes.
NBB : dans le cas où on a des valeurs aberrantes (*outliers*), on peut soit ajuster une « droite de résistance » basée sur la différence Y'- médiane_X (au lieu de $\overline{Y}_X$), soit tout simplement éliminer ces valeurs.

Propriétés de la régression :

- Toutes les droites ($\forall$r) passent par l'intersection de $\overline{X}$ et $\overline{Y}$ (pivot → on peut calculer r à partir d'<u>une</u> valeur de Y').
- Si r = 0, $Y' = \overline{Y}$.

Equation de la régression : $\frac{Y'-\overline{Y}}{S_Y} = r\,\frac{X-\overline{X}}{S_X} \rightarrow Y' = (r\,\frac{S_Y}{S_X})X - (r\,\frac{S_Y}{S_X})\,\overline{X} + \overline{Y}$
Méthode pratique de calcul :
$Y = aX + b$

$$a = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}$$

$b = \overline{Y} - a\,\overline{X}$

Erreur-type d'estimation de Y en X (mesure de la variablité de Y par rapport à Y') :

$$S_{YX} = \sqrt{\frac{\sum(Y-Y')^2}{n}} = S_Y\sqrt{1-r^2}$$

Si r = 1 → $S_{YX} = 0$; si r = 0 → $S_{YX} = S_Y$ (écart-type de Y)
Attention : ne pas confondre S_{YX} avec S_{xy}, covariance de X et Y.
L'utilisation légitime de S_{YX} dépend de 3 conditions :

- Linéarité X-Y
- Homoscédasticité (même variance de Y pour tous les X)
- Distribution normale de Y pour tous les X

→ Toujours inspecter les données brutes sur le graphe !
NB : la valeur de la prédiction dépend toujours de la taille de l'échantillon.
NBB : si on trace 2 lignes parallèles autour de la droite de régression à des distances de 1 S_{YX}, 2 S_{YX} ou 3S_{YX}, on devrait, si n est grand, trouver entre ces lignes respectivement 68%, 95% ou 99,7% des points du nuage.

Interprétation de r

$S_{YX} = S_Y\sqrt{1-r^2}$ → $r = \sqrt{1-\frac{S_{YX}^2}{S_Y^2}}$

→ r n'est pas seulement fonction de la dispersion de S_{YX} (ΔY par rapport à la régression), mais du rapport $\frac{S_{YX}}{S_Y}$.

Si $S_{YX} = 0$ (corrélation parfaite) → $r = \pm 1$
Si $S_{YX} = S_Y$ → $r = 0$.

Notion de « gamme de talents » :
Ex : stress ou pollution / taille des villes : faire Aspremont → NY plutôt que Nice, Bordeaux, Lille, etc…
Avec S_{YX} = constante (homoscédasticité), r est en proportion directe de S_Y.

Ex : $r = \sqrt{1-\frac{10}{50}}$ ~0,9 ; $r = \sqrt{1-\frac{10}{20}}$ ~0,7

→ Quand S_Y (proportionnel à la gamme de talents) augmente, r augmente.

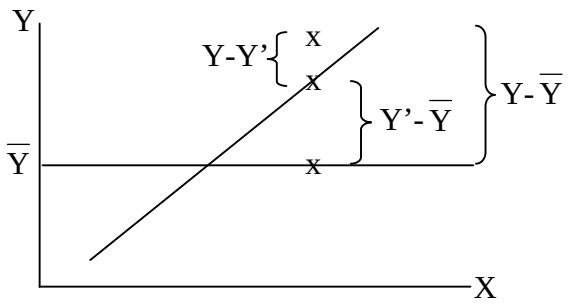
***

r est-il une indication de la pente d'une droite de régression ?
Y = aX + b où a = pente.

$Y' = \overbrace{(r\frac{S_Y}{S_X})}^{a}X + \overbrace{(-r\frac{S_Y}{S_X})\overline{X}+\overline{Y}}^{b}$ → pente $= r\frac{S_Y}{S_X}$ (appelé coefficient de régression)

→ si $S_Y = S_X$, r = pente ; mais dans la plupart des cas $S_Y \neq S_X$ → r indique de combien d'écarts-type Y augmente quand X augmente d'un écart-type ($z'_Y = r\, z_X$ en termes d'écarts réduits).

Coefficient de détermination



$\frac{\sum(Y-\overline{Y})^2}{n} = S_Y^2$ = variance totale de Y

$\frac{\sum(Y-Y')^2}{n} = S_{YX}^2$ = variance de Y indépendante de X (= résiduelle)

$\frac{\sum(Y'-\overline{Y})^2}{n} = S_{Y'}^2$ = variance de Y associée aux X (= expliquée)

$\frac{S_{Y'}^2}{S_Y^2}$ =… r^2 = coefficient de détermination → donne la proportion de la variance de Y associée avec le changement de valeur de X (corrélation).
NB : si r = 0,5 → $r^2 = 0,25 = 25\%$; si r = 0,71 → $r^2 = 0,5 = 50\%$…

Inférences sur la significativité du r

En prenant tous les échantillons possibles de taille n d'une population, on obtient une distribution d'échantillonnage des r dont la moyenne, ρ, est le vrai coefficient de corrélation de cette population. L'écart-type de cette distribution, $\sigma_r = \frac{1-\rho^2}{\sqrt{n-1}}$. On note que quand n ou ρ augmentent σ_r diminue.

Déterminer la significativité du r revient à tester l'hypothèse H_0: ρ = 0. Pour cela, on pourrait calculer $t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, et comparer à la valeur de t pour H_0 avec n-2 DL et $\alpha_{0,05}$ ou $\alpha_{0,01}$, directionnel ou pas.

Une méthode plus simple consiste à comparer r directement dans une Table de significativité avec DL = n-2.

Vu que σ_r diminue quand n augmente, on peut, avec des échantillons très grands, trouver un r hautement significatif alors que ρ est très faible. Inversement, avec un n très faible, on peut, par hasard, trouver un r élevé. Il vaudrait donc mieux donner l'intervalle de confiance de r pour une estimation plus… objective. Malheureusement, quand ρ≠0 la distribution de r n'est plus normale (elle l'est pour ρ=0, ce qui légitime le test H_0 ci-dessus).

→ Il faut convertir r en z_r de Fisher (attention : aucun rapport avec l'écart réduit z !) :

$z_r = ½ \ln\frac{1+r}{1-r}$ → distribution normale

La conversion est disponible dans une Table r → z_r.

Pour un IC de 95% : $z_r \pm 1{,}96\ \sigma_{z_r}$, où l'erreur-type de z_r, $\sigma_{z_r} = \frac{1}{\sqrt{n-3}}$.

On reconvertit ensuite les 2 limites z_r en valeurs de r.

***

De même, pour déterminer si une différence entre 2 r est significative, il faut convertir en z_r (valable pour des échantillons indépendants, plus complexe pour éch. dépendants).

$z = \frac{z_{r1}-z_{r2}}{\sigma_{z_{r1}-z_{r2}}}$, où l'erreur-type de la différence entre les 2 z_r, $\sigma_{z_{r1}-z_{r2}} = \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}$

→ Comparer ensuite à z = 1,96 pour α = 0,05 ou 2,58 pour α = 0,01.

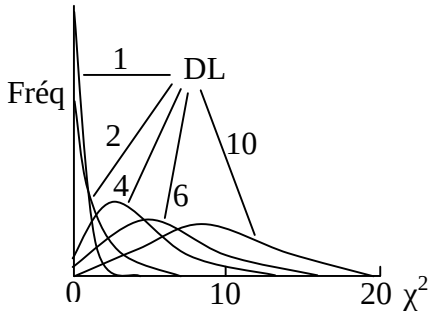
Quand on compare plus de 2 échantillons, il faut faire une analyse de covariance (ANCOVA ; voir par ex Scherrer : Biostatistique, 1984, p. 676). Ce type d'analyse permet de rechercher, parmi plusieurs droites de régression, une différence de pente et/ou d'ordonnée à l'origine.

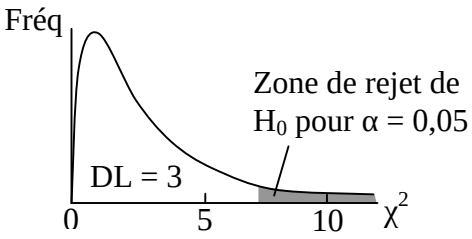
CHI-2 (χ^2 ; *chi-square*)

Permet d'estimer si les fréquences de distribution d'une population dans n catégories diffèrent des fréquences attendues selon une hypothèse quelconque. Ex : y-a-t'il une préférence des étudiants en LSV pour certaine(s) des 4 options proposées ?
On pourrait ici tester par ex. H_0: $p_A = p_B = p_C = p_D = 0{,}25$ (il n'y a pas de préférence).
Conséquemment, le χ^2 permet de tester la qualité ou la validité d'ajustement (*goodness of fit*) de données à une équation.

$\chi^2 = \sum^{n} \left(\frac{(f_o - f_a)^2}{f_a} \right)$ → Sommation, sur n catégories, des fréquences observées et attendues

(f_a = N x p = N x 0,25). On compare ensuite le χ^2 à la valeur critique dans une Table pour un α choisi et DL = nombre de catégories - 1. Si χ^2 > à la valeur critique, on rejette H_0.





Notas :
- Très employé par les généticiens qui attendent des fréquences particulières de reproduction (ex : apparition de phénotypes selon 9-3-3-1 → 9/16-3/16-3/16-1/16).
- Un prérequis est l'indépendance des observations. Pour utiliser le χ^2 en mesures répétées, il faut prendre certaines précautions (Siegel & Castellan, 1988).

Applications du χ^2

1) Ajustement d'équation
On a vu jusqu'ici qu'avec la plupart des tests on cherche une différence (rejet de H_0). Quand on utilise le χ^2 pour estimer la qualité d'ajustement, on cherche au contraire à démontrer qu'il n'y a pas de différence avec la distribution théorique (rétention de H_0).

$\chi^2 = \left(\frac{Y_1 - Y_1'}{\sigma_1} \right)^2 + \left(\frac{Y_2 - Y_2'}{\sigma_2} \right)^2 + \ldots + \left(\frac{Y_k - Y_k'}{\sigma_k} \right)^2$ où Y est la valeur observée pour un X donné,

Y' la valeur théorique déterminée par l'équation pour ce même X, et σ l' erreur-type de Y.
DL = nombre de données - nombre de paramètres variables

2) Tableaux de contingence à 2 facteurs : permettent de tester l'indépendance des groupes ou au contraire l'interaction entre 2 facteurs. NB : variables <u>qualitatives</u> surtout utilisées. Ex :

		PACA	IdF	Nord	
Leg +	f_o	2	2	10	14
	f_a	3,5	2,8	7,7	
Leg -	f_o	8	6	12	26
	f_a	6,5	5,2	14,3	
Prélèvements		10	8	22	

= 40 prélèvements

H_0: la proportion de contamination est identique dans toutes les régions.
$f_a = 14/40 = 0,35$ → $f_{aPACA} = 10 \times 0,35 = 3,5$; $f_{aIdF} = 8 \times 0,35 = 2,8$; $f_{aNord} = 22 \times 0,35 = 7,7$.

$\chi^2 = \frac{(2-3,5)^2}{3,5} + \frac{(8-6,5)^2}{6,5} \ldots = 2,4$. Avec DL = (col-1)(lignes-1) = 2, $\chi^2 = 6$ pour $\alpha = 0,05$.

→ On ne peut pas exclure que les prélèvements sont homogènes.

NB : quand χ^2 est significatif, on ne peut pas localiser l'hétérogénéité, c-à-d déterminer quels groupes sont dissemblables. Il faut pour ça faire une transformation log complexe…
NBB : le χ^2 a beaucoup d'autres applications possibles (voir plus bas par ex.).

STATISTIQUES NON-PARAMÉTRIQUES

Pour distributions non-normales.
Ne testent pas une différence entre moyennes (paramètre) mais entre distributions, et peuvent s'adresser indifféremment à la tendance centrale, dispersion ou symétrie.
Adaptées aux petits échantillons.
Basées sur un classement des données par rangs. Ex :

Valeurs :	4	5	5	8	11	11	11	15	19
Rang :	1	2,5	2,5	4	6	6	6	8	9
		moyennes		des		rangs			

I. 2 groupes indépendants : Mann-Whitney U (équivalent du t)
H_0: les 2 échantillons viennent de populations avec la même distribution.
NB : pour des formes et dispersions similaires, le test évalue surtout la tendance centrale, laquelle se rapproche plus de la médiane.

Procédure :
- Classer toutes les valeurs (X, Y) par rangs.
- Si $n_X < n_Y$, calculer ΣR_X, la somme des rangs pour X.
- Comparer ΣR_X dans une Table avec $\alpha = 0{,}025$ pour $p_{0,05}$ bidirectionnel, n_X et n_Y.

Si ΣR_X < au chiffre inférieur ou > au chiffre supérieur, rejeter H_0.
Pour $\alpha = 0{,}05$ directionnel, regarder seulement < ou >.
Quand n > aux n de la Table, il faut calculer U :

$$U = (n_X)(n_Y) + \frac{n_X(n_X+1)}{2} - \Sigma R_X$$

... puis l'écart réduit, $z = \dfrac{U-(n_X n_Y/2)}{\sqrt{\dfrac{n_X n_Y(n_X+n_Y+1)}{12}}}$ → > à 1,96 ou 2,58 ?…

Autre test possible : Kolmogorov-Smirnov

II. Plus de 2 groupes indépendants : Kruskal-Wallis (équivalent d'ANOVA)

$$H = -3\,(n_t+1) + \frac{12}{n_t(n_t+1)}\left(\frac{(\Sigma R_1)^2}{n_1}+\frac{(\Sigma R_2)^2}{n_2}+\ldots+\frac{(\Sigma R_k)^2}{n_k}\right)$$

On compare H à la valeur critique du χ^2_α avec DL = k-1.
Pour les comparaisons multiples (H > à la valeur critique), on utilise les tests post-hoc utilisés en stats paramétriques. Pour comparer tous les groupes entre eux, et si les n de chaque groupe sont égaux, on peut par ex. utiliser le test de Student-Newman-Keuls en remplaçant les moyennes par la somme des rangs. Avec des n inégaux et/ou pour comparer les groupes à un témoin, on peut utiliser les tests de Dunn ou Dunnett. Pour plus de détails, voir Biostatistique de Scherrer, p. 540.

III. 2 groupes dépendants

1) Test des signes : plus adapté à des variables semi-quantitatives (échelles ordinales). Pour chaque paire, on note la différence de résultat en tant que signe + ou -.
   H_0: il y a autant de + que de -, ce qui équivaut à un test de fréquences. On peut donc utiliser le χ^2.

$\chi^2 = \frac{(f_{o+}-f_{a+})^2}{f_{a+}} + \frac{(f_{o-}-f_{a-})^2}{f_{a-}}$ … et voir Table…

NB : quand une différence = 0, éliminer la paire et réduire N (pour le calcul de f) en conséquence.

NBB : le test des signes est OK pour n ≥ 10 (sinon test binômial avec les précautions indiquées par ex. par Scherrer, p. 524).

2) Wilcoxon : le test de choix pour des variables quantitatives non-normales.
Plus puissant que le test des signes car il tient compte de la taille des différences entre paires.
On classe par rangs la valeur absolue de la différence X-Y, puis on lui attribue le signe de la différence. On fait la somme (en valeurs absolues) des valeurs positives (W_+) et négatives (W_-). Ex :

X	Y	X-Y	Rang	Signe
24	28	4	1	-1
39	29	10	3	+3
29	34	5	2	-2

$W_+ = 3$; $W_- = 3$

On prend la valeur la plus petite (W_+ ou W_-) qu'on compare dans une Table selon α et le nombre de paires.

Pour les n non-inclus dans la Table : $z = \frac{W - 0{,}25n(n+1)}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$ où 0,25n(n+1) est la moyenne de la distribution des W.

IV. Plus de 2 groupes dépendants : faire une ANOVA de Friedman (voir fichier d'aide de Statistica).

Coefficient de corrélation de Spearman (r_s)

A utiliser par ex. quand n est petit.

$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$, où D = différence entre les paires de rangs et n = nb de paires de rangs

Ici on classe chaque groupe X et Y par rangs. Ex :

X	Y	Rang X	Rang Y	D	D^2
1	5	1	2	-1	1
3	4	2	1	1	1
5	20	3	5	-2	4
7	15	4	3	1	1
9	19	5	4	1	1

$r_s = 1 - \frac{6 \times 8}{5(25-1)} = 0{,}6$ → on compare ensuite r_s à la valeur appropriée pour α et DL = n-2 dans la même Table que pour le r de Pearson (0,6 < 0,88 → H_0 est retenue).