LES TESTS STATISTIQUES HABITUELS

Bernard LEGRAS

Petit extrait de ses ouvrages de statistique

SOMMAIRE

I Le principe et les risques d'un test

I-1 Exemple

On considère une variété de souris qui présente des cancers spontanés avec une fréquence moyenne parfaitement connue f_0 = 20 %. On expérimente un produit dans le but de savoir s'il agit sur ce taux.
Pour cela, on procède à une *expérience sur un échantillon de 100 souris* et on obtient une fréquence de cancers égale à f. Le problème est de savoir si le produit est actif ou non.

On commence par choisir une *hypothèse* qui permette le test.
Ici pour répondre à cette question, *on suppose que le traitement est sans effets*, c'est-à-dire que l'échantillon est extrait de la population générale caractérisée par f_0.
C'est *l'hypothèse à tester ou hypothèse* H_0.

Dans ce cas, les fluctuations d'échantillonnage font varier f autour de 20 % et on peut les limiter à un intervalle avec une probabilité donnée.
Nous savons par exemple qu'au risque 5 % correspond un écart à la moyenne d'environ 2 σ (approximation normale). On a ici (loi binomiale) :

$\sigma^2 = f_0 (1 - f_0) / n = 0{,}2 \times 0{,}8 / 100 = 0{,}0016$ soit $\sigma = 0{,}04$

Ainsi f n'a que 5 chances sur 100 de sortir de l'intervalle $20 \pm 2 \times 4$ % soit 12 à 28 %.

On adopte alors la *règle* suivante :

- si le pourcentage observé f tombe à l'intérieur de l'intervalle I = 12 - 28 %, nous admettrons que l'activité du traitement n'est pas prouvée.
- par contre, si f est en dehors de l'intervalle, nous n'admettrons plus que l'écart à 20 % résulte des fluctuations d'échantillonnage ; nous dirons qu'il est *significatif* et le traitement sera considéré comme actif.

I-2 Risques

Le test de signification comporte *2 risques d'erreur.*

a) *Le traitement est inactif*
Alors f fluctue autour de 20 %, mais il peut arriver que les fluctuations d'échantillonnage l'amènent en dehors de l'intervalle ("mauvais" échantillon).
Comme dans ce cas, nous déclarons le traitement efficace, nous commettons une erreur : ce risque d'erreur est parfaitement connu : c'est ici 5 %.

b) *Le traitement est actif*
Dans ce cas f fluctue autour d'une valeur f'_0 différente de 20 %, mais il peut arriver que les fluctuations d'échantillonnage amènent f dans l'intervalle 12 - 28 % (surtout si f'_0 est proche de f_0).

Dans cette éventualité, nous déclarons que l'activité du traitement n'est pas prouvée et nous écartons à tort un produit efficace ; on dit que c'est un *manque de puissance du test.*

Ainsi le test comporte *2 risques* :

α = *risque de première espèce*. C'est la probabilité de mettre en évidence une différence qui n'existe pas.

ß = risque de seconde espèce. C'est la probabilité de ne pas mettre en évidence une différence qui existe réellement.

On préfère en général travailler sur le complément de ß.

1 - ß = puissance du test. C'est la probabilité de mettre en évidence une différence qui existe réellement.

Risque de 2ème espèce : il dépend de *plusieurs facteurs* :

α (antagonisme entre les 2 risques)

n

$f_1 - f_2$ pour la comparaison de 2 fréquences

ou $\mu_1 - \mu_2$ et σ pour la comparaison de 2 moyennes.

On peut évaluer les modifications dues à ces paramètres à l'aide de graphiques simples.

Intérêt pratique du risque ß : avant tout, la *détermination du nombre de sujets nécessaires* pour obtenir un risque ß fixé (enquête, essai...). On peut utiliser des formules ou des tables.

I-3 Hypothèse nulle ; hypothèse alternative

L'énoncé du problème tel que nous l'avons fait jusqu'ici est en réalité insuffisant. En effet, il convient en outre de préciser l'hypothèse que nous retiendrons si l'on rejette H_0. Cette hypothèse dite *alternative est appelée* H_1.

Sa connaissance influe sur le test : bilatéral ou unilatéral.

On distingue *2 possibilités* :

- 1er cas (le plus général)

On veut savoir s'il existe une différence *quel que soit son signe.*

Ex : comparaison de 2 moyennes.

Les 2 échantillons sont extraits de 2 populations de moyenne μ_1 et μ_2.

H_0 $\mu_1 = \mu_2$

H_1 $\mu_1 \neq \mu_2$ soit $\mu_1 > \mu_2$

ou $\mu_1 < \mu_2$

On pratique alors un *test bilatéral*.

On peut remarquer que c'est ce qui a été réalisé avec l'exemple des souris. On définit une zone de rejet de H_0 de part et d'autre de l'intervalle d'acceptation.

- 2ème cas (moins fréquent)

On veut savoir s'il existe une différence *dans un sens donné*.

Ex : comparaison de 2 moyennes

H_0 $\mu_1 = \mu_2$

H_1 $\mu_1 < \mu_2$

Dans cette situation, on est conduit à un *test unilatéral* avec une zone de rejet de H_0 d'un seul côté.

II Tests sur les moyennes

II-1 Résultats fondamentaux

Considérons une variable X de moyenne μ et d'écart-type σ.
Nous n'envisagerons que le cas habituel où σ est inconnu.

σ inconnu

On suppose que X est normal (en pratique, il faudrait le vérifier).
On considère la variable centrée et réduite suivante :

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n-1}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n-1}}}$$

Question : quelle est sa loi ?

Nous allons démontrer qu'il s'agit d'une *loi de Student à n- 1 degrés de liberté (ddl)*

Pour cela, divisons le numérateur et le dénominateur par $\sqrt{\frac{\sigma^2}{n}}$

En réalité pour le dénominateur, on multipliera par l'inverse soit $\sqrt{\frac{n}{\sigma^2}}$

On obtient :

$$\frac{\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{\frac{nS^2}{\sigma^2}}{n-1}}} = \frac{U}{\sqrt{\frac{X^2_{n-1}}{n-1}}}$$

Nous avons signalé que $\frac{nS^2}{\sigma^2}$ suit une loi du X^2 à n - 1 ddl si X est normale.

Donc, loi de probabilité = loi de Student à n - 1 ddl.

On aboutit finalement au résultat suivant :

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n-1}}} = T_{n-1}$$

II-2 Intervalle de confiance de la moyenne

Le problème est facile à résoudre à partir des résultats que nous venons de démontrer.

σ inconnu
(X est normal)

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n-1}}} = T_{n-1}$$

Comme précédemment, on se fixe le risque α.
La table de Student donne le seuil t_α correspondant au risque α et à n - 1 degré de liberté.
On peut alors écrire :

$$1 - \alpha = Pr\ (- t_\alpha < T_{n-1} < t_\alpha\ )$$

Revenons à $\overline{X}$ en remplaçant T_{n-1} par sa valeur :

$$1 - \alpha = Pr\ (- t_\alpha < \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n-1}}} < t_\alpha)$$

$$= Pr\ (\overline{X} - t_\alpha \frac{S}{\sqrt{n-1}} < \mu < \overline{X} + t_\alpha \frac{S}{\sqrt{n-1}}\ )$$

L'intervalle $IC(1-\alpha) = \overline{X} \pm t_\alpha \frac{S}{\sqrt{n-1}}$ est l'intervalle de confiance de μ au risque α.

Remarque : en pratique, on a un échantillon de moyenne $\bar{x}$ et d'écart-type s.
On appellera aussi intervalle de confiance de la moyenne μ :

$$IC(1-\alpha) = \bar{x} \pm t_\alpha \frac{s}{\sqrt{n-1}}$$

II-3 Comparaison à une moyenne théorique

Hypothèse H_0 : μ_0 est la moyenne véritable de la population.

σ inconnu
(X est normal)

On calcule ici l'écart-réduit suivant : $t_{cal} = (\bar{x} - \mu_0) \Big/ \frac{s}{\sqrt{n-1}}$

Comme nous l'avons vu, c'est une réalisation d'une variable de Student à n-1 ddl.

L'intervalle de confiance de T_{n-1} est $(-t_\alpha, t_\alpha)$.

On compare donc $|\ t_{cal}\ |$ à t_α donné dans les tables.

Règle de décision :

si $|t_{cal}| < t_\alpha$ H_0 acceptée

sinon H_0 rejetée

En pratique $\alpha = 0{,}05$

si $|t_{cal}| < t_{0,05}$ différence non significative

sinon différence significative

c) Degré de signification

Définition : *c'est le risque d'erreur (P) correspondant à la valeur du paramètre calculé sous l'hypothèse H_0.*

Ex : si on se ramène à une loi normale, on calcule u_{cal}

$$\text{---> } P = Pr\,(|U|) > u_{cal})$$

En général, on fournit une *valeur approchée indiquant l'ordre de grandeur* (on ne fait pas d'interpolation linéaire). L'habitude est d'indiquer le degré de signification *seulement quand on rejette H_0,* c'est-à-dire quand $P < \alpha$.

II-4 Comparaison de 2 moyennes : échantillons indépendants

a) Généralités

Soient 2 échantillons indépendants définis par :

n_1 $\bar{x}_1$ s_1

n_2 $\bar{x}_2$ s_2

Le problème est le suivant :

les 2 moyennes $\bar{x}_1$ et $\bar{x}_2$ diffèrent-elles significativement ou non ?

On peut schématiser le problème de la façon suivante :

- l'échantillon 1 est tiré d'une population 1 où la variable X_1 a une moyenne μ_1 et un écart-type σ_1.

- l'échantillon 2 est tiré d'une population 2 où la variable X_2 a une moyenne μ_2 et un écart-type σ_2.

Nous allons tester l'hypothèse suivante :

$H_0 : \mu_1 = \mu_2$

$\bar{x}_1$ et $\bar{x}_2$ sont des réalisations de 2 variables $\overline{X}_1$ et $\overline{X}_2$.

Nous supposerons $\overline{X}_1$ et $\overline{X}_2$ normales (n grand ou X normale si n est petit).

On considère la variable $D = \overline{X}_1 - \overline{X}_2$.

Propriétés de D :

- $E(D) = E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2 = 0$ (sous H_0)

- $Var(D) = Var(\overline{X}_1 - \overline{X}_2) = Var(\overline{X}_1) + Var(\overline{X}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

(additivité des variances car échantillons indépendants)

- loi D = loi normale (différence de variables normales).

On en déduit que : $\frac{D-E(D)}{\sigma_D} = U$

soit $\frac{D}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = U$

b) σ_1 et σ_2 inconnus - test t de Student-Fisher

Conditions :

- les échantillons doivent être gaussiens.
- la solution n'est simple mathématiquement que si les variances σ_1^2 et σ_2^2 sont égales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). Il faut donc que les écarts-types s_1 et s_2 ne soient pas différents significativement (on peut le vérifier par le test F décrit plus loin).
- le test dit test t de Student-Fisher est valable quel que soit n.

On aboutit au résultat suivant :

Sous l'hypothèse H_0 :

$$\frac{D}{\sigma_D} = \frac{D}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\right)}} = T_{n_1+n_2-2}$$

C'est une variable qui suit une loi de Student à n_1+n_2-2 ddl.
Il suffit donc de calculer :

$$t_{cal} = \frac{d}{s_d} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right)}}$$

et de comparer $|t_{cal}|$ à t_α donné par la table de Student pour α et n_1+n_2-2 ddl .
Règle de décision :
si $|t_{cal}| < t_\alpha$ H_0 acceptée
sinon H_0 rejetée.
En pratique $\alpha = 0,05$
si $|t_{cal}| < t_{0,05}$ différence non significative entre $\bar{x}_1$ et $\bar{x}_2$.
sinon différence significative et on cherche le degré de signification

II-5 Valeurs appariées

Ex : Valeurs obtenues avant et après traitement *chez les mêmes individus.*
Dans ce cas, les échantillons *ne sont pas indépendants*, et les formules obtenues ne sont plus valables (non additivité des variances).

La méthode utilisée consiste à déterminer pour chaque couple de valeurs (x_1 , x_2) la différence $d = x_1 - x_2$.

- L'hypothèse H_0 est que les x_1, x_2 proviennent d'une même population.

C'est-à-dire : la moyenne véritable des différences est nulle
$E(D) = 0$ avec $D = X_1 - X_2$ (formule différente de $D = \overline{X}_1 - \overline{X}_2$ pour échantillons indépendants).

Le test est donc un cas particulier du test de comparaison d'une moyenne théorique μ_0 à une moyenne observée $\bar{x}$ avec $\mu_0 = 0$ et $\bar{x} = \bar{d}$
Nous n'envisagerons ici que le cas où σ est inconnu.

Test = *test t de Student-Fisher sur valeurs appariées.*
Comme nous l'avons vu :

$$t_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}} \quad \text{soit dans ce cas} \quad \frac{\bar{d}}{\frac{s_d}{\sqrt{n-1}}}$$

n = nombre de couples de valeurs
$\bar{d}$ = moyenne des différences et s_d = écart-type des différences

t_{cal} = réalisation de T_{n-1} (si les différences suivent la loi normale).
On compare $|t_{cal}|$ à t_α avec n - 1 ddl. Règle de décision habituelle.

II-6 Nombre de sujets nécessaires

On démontre un certain nombre de formules fournissant n_1 et n_2 en fonction de α, ß, Δϕ ou Δμ et σ.

En pratique, il suffit seulement de savoir utiliser les tables.

5 tables
- moyenne | test unilatéral avec α = ß = 5 %
- pourcentage | test bilatéral avec α = ß = 5 %
- facteur de correction quand α et (ou) ß ≠ 5 % (test uni ou bilatéral)

Elles correspondent au cas où $n_1 = n_2 = n$.
Elles donnent le nombre de sujets minimal dans chaque groupe en fonction des caractéristiques :

Δ et σ	test sur les moyennes
P et P' (%)	test sur les fréquences.

III Tests sur les fréquences

III-1 Rappel des résultats fondamentaux

Soit une population où les individus présentent le caractère A avec une fréquence f_0. Soit un échantillon d'effectif n tiré de la population (échantillon représentatif). A y est obtenu avec une fréquence f.
f est une réalisation d'une variable F dont nous avons vu les propriétés (lors de l'étude de la loi binomiale).

Propriétés de F :
- loi de F = loi binomiale
- la loi de F → loi normale.

La convergence est d'autant plus rapide que f_0 est voisin de 0,5.
En pratique, l'approximation est en général acceptable lorsque
$n f_0$ et $n (1 - f_0) > 10$ (si f_0 compris entre 0,1 et 0,9)
- $E(F) = f_0$
- $Var(F) = f_0(1 - f_0)/n$

Dans les différents tests, nous *supposerons que l'on peut faire l'approximation normale*. Vous verrez alors que les tests sont simples et très comparables à ceux relatifs aux moyennes. Ils sont en réalité moins rigoureux à cause de l'approximation normale.

III-2 Intervalle de confiance d'une fréquence

La variable centrée et réduite est (approximation normale) :

$$\frac{F-E(F)}{\sigma_F} = \frac{F-f_0}{\sqrt{\frac{f_0(1-f_0)}{n}}} \approx U$$

On se fixe le risque α ; la table P(u) donne le seuil u_α correspondant.
On a $1 - \alpha = Pr(-u_\alpha < U < u_\alpha)$.
Soit en revenant à F :

$$1 - \alpha \approx Pr\left(-u_\alpha < \frac{F-f_0}{\sqrt{\frac{f_0(1-f_0)}{n}}} < u_\alpha\right)$$

$$1 - \alpha \approx Pr\left(F - u_\alpha \sqrt{\frac{f_0(1-f_0)}{n}} < f_0 < F + u_\alpha \sqrt{\frac{f_0(1-f_0)}{n}}\right)$$

L'intervalle $IC(1-\alpha) = F \pm u_\alpha \sqrt{\frac{f_0(1-f_0)}{n}}$ est l'intervalle de confiance de f_0 au risque α. Il y a une probabilité $\approx 1 - \alpha$ pour que f_0 soit à l'intérieur de cet intervalle.

En pratique, on a un échantillon et une réalisation f de f_0. Bien que cela soit un peu moins correct, on appellera aussi intervalle de confiance de la fréquence ;

$$IC(1-\alpha) = f \pm u_\alpha \sqrt{\frac{f\,(1-f)}{n}}$$

III-3 Comparaison à une fréquence théorique

Hypothèse H_0 : f_0 = fréquence de la population
(Approximation normale).

$$\frac{F-E(F)}{\sigma_F} = \frac{F-f_0}{\sqrt{\frac{f_0(1-f_0)}{n}}} \approx U$$

Il suffit donc de calculer :

$$\left| u_{cal} \right| = \frac{\left| f-f_0 \right|}{\sqrt{\frac{f_0(1-f_0)}{n}}}$$ et de le comparer à u_α.

Règle de décision :

si $\left| u_{cal} \right| < u_\alpha$ H_0 acceptée

sinon H_0 rejetée

En pratique $\alpha = 0{,}05$

si $\left| u_{cal} \right| < 2$ différence non significative entre f et f_0

sinon différence significative et on cherche le degré de signification.

III-4 Comparaison de 2 fréquences

Soient : 2 échantillons *indépendants* définis par :

n_1 f_1

n_2 f_2

Comme pour les moyennes, il s'agit de savoir si f_1 et f_2 diffèrent de façon significative ou pas.

Pour cela on teste l'hypothèse H_0 suivante : les 2 échantillons sont extraits d'une même population qui présente une fréquence f_0 pour le caractère étudié.

f_1 et f_2 sont des réalisations de 2 variables binomiales F_1 et F_2.

On suppose l'approximation normale.

On considère la variable $D = F_1 - F_2$.

Propriétés de D :

- loi de D ≈ loi normale (différence de variables normales)
- $E(D) = E(F_1 - F_2) = E(F_1) - E(F_2) = f_0 - f_0 = 0$ (sous H_0)
- $Var(D) = Var(F_1 - F_2) = Var(F_1) + Var(F_2)$ (échantillons indépendants)

$$= f_0 (1 - f_0) / n_1 + f_0 (1 - f_0) / n_2$$

$$= f_0 (1 - f_0) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

On forme la variable réduite :

$$\frac{D - E(D)}{\sigma_D} = \frac{D}{\sqrt{f_0 (1 - f_0) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = U$$

Mais f_0 n'est en général pas connu. Il faut remplacer f_0 par sa meilleure estimation.

On montre que cette *meilleure estimation* s'obtient en considérant que l'on est en présence d'*un seul échantillon* d'effectif $n_1 + n_2$.

Le nombre d'individus possédant le caractère étudié est $n_1 f_1 + n_2 f_2$ et par suite la fréquence estimée vaut :

$$\hat{f}_0 = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

soit : $\dfrac{\text{nombre d'individus possédant le caractère étudié}}{\text{nombre total d'individus}}$

On a alors :

$$\frac{D}{\sqrt{\hat{f}_0 (1 - \hat{f}_0) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = U$$

Le test est alors facile. On calcule :

$$|u_{cal}| = \frac{|f_1 - f_2|}{\sqrt{\hat{f}_0 (1 - \hat{f}_0) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

et on le compare à u_α.

Règle de décision :

si $|u_{cal}| < u_\alpha$ H_0 acceptée

sinon H_0 rejetée

En pratique $\alpha = 0{,}05$:

si $|u_{cal}| < 2$ différence non significative entre les 2 fréquences

sinon différence significative et on cherche le degré de signification

IV Tests sur les variances

IV-1 Intervalle de confiance de la variance (échantillon normal)

Rappel : si X est normal, alors $\frac{nS^2}{\sigma^2}$ suit X^2_{n-1}

On choisit le risque α. Il lui correspond les valeurs seuil X_1^2 et X_2^2 .

On peut écrire : $1 - \alpha = Pr\ (X_1^2 < X^2 < X_2^2)$

Remplaçons $1 - \alpha = Pr\ (X_1^2 < \frac{nS^2}{\sigma^2} < X_2^2)$

soit : $= Pr\left(\frac{nS^2}{X_2^2} < \sigma^2 < \frac{nS^2}{X_1^2}\right)$

Intervalle de confiance de σ^2 : $IC(1-\alpha) = \left(\frac{nS^2}{X_2^2} : \frac{nS^2}{X_1^2}\right)$

Si l'on a un échantillon de variance s^2, alors :

$$IC(1-\alpha) = \left(\frac{ns^2}{X_2^2} : \frac{ns^2}{X_1^2}\right)$$

IV-2 Loi de Snedecor

Le test habituel de comparaison de 2 variances est basé non pas sur leur différence (comme dans le cas des moyennes et des fréquences) mais sur le rapport des 2 variances estimées.
Ce test est dit *test F ou test de Snedecor* (car on se ramène à une loi de Snedecor).
Avant de l'envisager, il faut étudier la loi de Snedecor.

a) Résultats de base

Soit une population normale de variance σ^2.
On extrait de cette population 2 échantillons indépendants définis par :

n_a s_a^2

n_b s_b^2

Snedecor a étudié la variable suivante (rapport des variances estimées)

$$F = \frac{\frac{n_a S_a^2}{n_a - 1}}{\frac{n_b S_b^2}{n_b - 1}}$$

Elle s'écrit aussi en divisant au numérateur et au dénominateur par σ^2.

$$\frac{\frac{n_a S_a^2}{\sigma^2}}{\frac{n_a - 1}{\frac{n_b S_b^2}{\sigma^2}}} = \frac{\frac{\chi^2_{n_a-1}}{n_a - 1}}{\frac{\chi^2_{n_b-1}}{n_b - 1}}$$

(variable de Snedecor)

On sait calculer sa densité de probabilité (formule compliquée).
La courbe représentative est dissymétrique, de moyenne légèrement supérieure à 1.
La loi de probabilité de F est donc connue et peut être résumée dans les tables.
Ces tables permettent de déterminer les limites f_α et f'_α pour différents risques α en fonction des ddl : $l_a = n_a - 1$ et $l_b = n_b - 1$

Au risque α, la variable F est comprise entre les valeurs seuil f'_α et f_α
--> intervalle de confiance de F = (f'_α, f_α) au risque α choisi.

f_α valeur supérieure (> 1) f'_α valeur inférieure (< 1)

Remarque :

- ces valeurs dépendent de *3 paramètres* ;

Comme on ne peut constituer que des tableaux à 2 entrées, on se fixe la probabilité et on fait varier les ddl.

- par ailleurs, il faudrait théoriquement, pour chaque valeur, 2 tables donnant l'une f_α et l'autre f'_α.

- nous allons voir que l'on peut se limiter à la table donnant directement f_α et qu'on obtient f'_α par un petit calcul simple.

b) Tables de Snedecor

En pratique : les tables donnant f_α correspondent à la probabilité P telle que $P = Pr(F > f_\alpha)$ avec F > 1. On constate donc que $P = \alpha / 2$.

Utilisation des tables :

1 - *limite supérieure f_α (> 1)*

On considère la variance estimée la plus élevée (pour avoir F > 1).

Ex :

$$\frac{n_a s_a^2}{n_a-1} > \frac{n_b s_b^2}{n_b-1} \qquad F = \frac{\frac{n_a s_a^2}{n_a-1}}{\frac{n_b s_b^2}{n_b-1}} > 1$$

On prend la table P = α / 2 si le risque est α.
La borne supérieure f_α est lue dans la table à l'intersection des ddl :
$l_a = n_a - 1$ définit la colonne et $l_b = n_b - 1$ définit la ligne.

2 - limite inférieure f' α (< 1) (nous admettons le résultat)

On permute les ddl pour calculer la limite inférieure.
Dans la même table correspondant au risque $\alpha/2$, à l'intersection de l_b et l_a , on démontre qu'on obtient $\frac{1}{f'_\alpha}$ d'où f'_α.

Remarque : f_α et f'_α sont des bornes inégalement distantes de 1 du fait de l'asymétrie de la courbe.

IV-3 Comparaison de 2 variances par le test de Snedecor

Les 2 échantillons sont extraits de 2 populations normales de variances σ_a^2 et σ_b^2 .
On teste l'hypothèse $H_0 : \sigma_a^2 = \sigma_b^2$
On se fixe un risque α (par ex 5 %).

On calcule $f_{cal} = \frac{\frac{n_a s_a^2}{n_a-1}}{\frac{n_b s_b^2}{n_b-1}}$

Attention : on s'arrange pour mettre au numérateur la valeur la plus élevée pour avoir $f_{cal} > 1$.
f_{cal} = réalisation de la variable F. On la compare aux valeurs seuils.
Il est alors inutile de comparer f_{cal} à la borne inférieure (< 1).
Il suffit de comparer f_{cal} à la borne supérieure f_α correspondant à la probabilité $\alpha/2$.
Règle de décision :
si $f_{cal} < f_\alpha$ on accepte H_0 : les 2 variances ne diffèrent pas significativement
sinon on rejette H_0 : les différences sont significatives au risque choisi.

V Le test du Chi deux ($X^{2)}$

Ce test est utilisé :
- pour comparer une distribution théorique avec une distribution expérimentale
- pour tester l'indépendance de 2 variables (cas particulier du cas précédent).

V-1 Test de conformité

a) Méthode générale

Elle consiste à comparer des valeurs observées (O) à des valeurs théoriques (T).
Hypothèse H_0 : O et T proviennent d'une même population.
Le choix de T repose :
- soit sur des raisons théoriques
- soit sur des résultats expérimentaux

b) Relation fondamentale

Soit une distribution statistique constituée de n observations rangées en k classes (ce rangement en classes est en effet indispensable). Considérons la classe d'indice i ($i \leq k$).
Soient : k_i l'effectif de la classe i.
p_i la probabilité théorique.
$n\,p_i$ l'effectif théorique.

Remarque : nous appelons k_i l'effectif mesuré plutôt que n_i pour bien le distinguer de $n\,p_i$ l'effectif théorique.
Pour caractériser la différence entre l'effectif théorique et l'effectif réel de la classe i, on étudie l'écart suivant appelé *écart quadratique relatif* :

$$\frac{(k_i - n\,p_i)^2}{n\,p_i}$$

Pour l'ensemble de la distribution, on définit le paramètre positif X^2_{cal} qui est la somme de ces écarts pour les k classes.

$$X^2_{cal} = \sum_{i=1}^{k} \frac{(k_i - n\,p_i)^2}{n\,p_i}$$

Si les écarts sont faibles pour chacun de ces termes, la valeur totale sera peu élevée.
Si au contraire, les écarts sont grands, la valeur totale sera importante.

Voyons maintenant le lien avec la loi du X^2.
PEARSON a étudié la variable :

$$Y = \sum_{i=1}^{k} \frac{(K_i - n\,p_i)^2}{n\,p_i}$$

K_i (majuscule) est la variable qui correspond à la réalisation k_i.
$n\,p_i$ est écrit en minuscule car il s'agit d'une valeur définie.
Sous l'hypothèse H_0, on démontre (démonstration complexe) :

quand $n \rightarrow \infty$ $\quad Y \rightarrow U_1^2 + U_2^2 + \ldots\ldots + U_{k-1}^2$ (avec U_i indépendants).
Donc :

loi de $Y \rightarrow$ loi du X^2 à k-1 degrés de liberté quand $n \rightarrow \infty$

C'est pour cette raison que l'on note le paramètre X_{cal}^2 .
L'approximation est valable en pratique quand n > 30.

c) Degré de liberté

On peut se demander pourquoi l'on a k - 1 ddl alors que X_{cal}^2 est calculé sur k classes. La raison est qu'il y a bien k différences du type $k_i - n\,p_i$, mais il y en a *seulement k - 1 indépendantes.*
En effet, si k - 1 effectifs sont calculés, le k ème effectif est déterminé par la différence suivante : $n\,p_k = n - (n\,p_1 + n\,p_2 + \ldots\ldots + n\,p_{k-1})$
Conclusion : il n'est pas indépendant et il y a seulement k - 1 effectifs indépendants.

d) Tables

La table habituelle donne les valeurs limites X_α^2 qui ont une probabilité α d'être dépassées.

$\alpha = Pr\,(X^2 \geq X_\alpha^2)$ α valeur choisie.

X_α^2 dépend de α et de l

e) Test

On fait l'hypothèse H_0 : distribution expérimentale et distribution théorique extraites d'une même population.
On calcule le paramètre :

$$X_{cal}^2 = \sum_{i=1}^{k} \frac{(k_i - np_i)^2}{n\,p_i}$$

C'est la réalisation d'une variable qui suit une loi du X^2 à k - 1 ddl (si n > 30).
On se fixe α, on lit X_α^2 dans les tables du X^2.
Règle de décision :
si $X_{cal}^2 < X_\alpha^2$ H_0 acceptée (distribution théorique acceptable).
sinon H_0 rejetée.

En pratique $\alpha = 0{,}05$:
si $X_{cal}^2 < X_{0.05}^2$ différences non significatives entre les 2 distributions

sinon différences significatives et on cherche le degré de signification.

V-2 Test d'indépendance

a) Cas de 2 variables à 2 modalités

Nous envisageons d'abord le cas le plus simple de 2 variables A et B à 2 modalités.
On veut savoir si la variable A est liée ou non à la variable B.
Le problème se traite en constituant le tableau de contingence suivant :

	A	$\bar{\mathbf{A}}$	effectif marginal
B	a	b	n_1
$\bar{\mathbf{B}}$	c	d	n_2
effectif marginal	n_3	n_4	n

Nous supposerons les effectifs élevés.
Hypothèse H_0 : les variables A et B sont indépendantes.

Sous cette hypothèse d'indépendance : la probabilité d'avoir un individu appartenant à une case (par ex (AB)) est donc d'après le théorème des probabilités composées : P (AB) = P (A) x P (B)
Mais les meilleures estimations des probabilités sont données par les fréquences.

$$p(A) = \frac{n_3}{n} \qquad p(B) = \frac{n_1}{n} \qquad \text{d'où : } p(AB) = \frac{n_3\ n_1}{n^2}$$

et l'effectif théorique correspondant vaut $n \cdot p(AB) = n \dfrac{n_3\ n_1}{n^2}$

$$n(AB) = \frac{n_3\ n_1}{n}$$

Conclusion : en pratique, il suffit donc de multiplier le total de la ligne par le total de la colonne et de diviser par le total général :

$$\text{effectif théorique} = \frac{\text{total ligne} \cdot \text{total colonne}}{\text{total général}}$$

On applique ensuite à ces 2 distributions (l'expérimentale et la théorique) la relation du X^2 et on calcule X^2_{cal} .

Un point important est le ddl. *Il est égal à 1 dans le cas des tableaux 2 x 2.*
En effet, une fois l'effectif théorique d'une case fixée, les autres s'en déduisent directement puisque l'on connaît les sommes de chaque ligne et de chaque colonne (1 seul effectif théorique indépendant).

On compare donc X^2_{cal} à X^2_α (α, 1) .

Règle de décision habituelle :

si $X^2_{cal} < X^2_\alpha$ H_0 acceptée

sinon H_0 rejetée.

b) Cas de 2 variables à plusieurs modalités

On calcule comme précédemment les effectifs théoriques et la valeur du X^2_{cal} . Les différences avec le cas précédent sont les suivantes :

- les différences entre effectif mesuré et effectif théorique ne sont plus les mêmes dans les différentes cases.
- surtout le nombre de degré de liberté est égal à :

$$l = (L - 1)\,(C - 1)$$

L = nombre de lignes et C = nombre de colonnes.

En effet, par ligne, on n'a que L - 1 effectifs indépendants puisque la somme par ligne est connue. De même par colonne ---> (C - 1). D'où le résultat.

V-3 Conditions d'utilisation

a) Conditions de validité du test

Pour appliquer le test du X^2, il faut :

- effectif total (n) au moins égal à 30.
- tous les effectifs théoriques (les np_i) doivent être > 5

(sinon regroupements si possible).

b) Correction de Yates

Il se pose un problème particulier pour les tableaux 2 x 2. En effet, la loi du X^2 est une loi continue, alors que le test du X^2 introduit des valeurs entières discontinues (k_i). L'erreur faite est très faible sauf lorsque le nombre de ddl est égal à 1 .

Dans ce cas, Yates a montré que c'est la variable

$$\sum_{i=1}^{k} \frac{(|K_i - n\,p_i| - 0{,}5)^2}{n\,p_i}$$ qui suit le mieux la loi du X^2.

La correction consiste donc à déterminer le paramètre corrigé suivant :

$$X^2_{cal} = \sum_{i=1}^{k} \frac{(|k_i - n\,p_i| - 0{,}5)^2}{n\,p_i}$$

Cela revient à diminuer la valeur absolue de chaque écart de la valeur 0,5.

Remarque : la correction est notable principalement pour n inférieur à 60.

Elle devient faible au-delà et peut alors ne pas être réalisée.

V-4 Cas des valeurs appariées (test de Mac Nemar)

Soit un essai portant sur n malades qui ont été leurs propres témoins : mesures avant et après. Ce résultat est noté + ou - .
On a donc 4 catégories de réponses :

Avant	Après	Effectif
-	-	n_1
-	+	a
+	-	b
+	+	n_2
		n

On ne peut pas utiliser les formules habituelles pour comparer les 2 traitements car ils ne sont pas indépendants.

Formules habituelles : % de + avant = $\frac{b+n_2}{n}$ après = $\frac{a+n_2}{n}$

On montre qu'il faut *laisser de côté les réponses concordantes.*
On examine seulement les effectifs - + (a) et + - (b).

Test du X^2

Sous l'hypothèse H_0 :

distribution observée	a	b
distribution théorique	$\frac{a+b}{2}$	$\frac{a+b}{2}$

on a 2 classes (k = 2) ddl = k - 1 = 1.

$$X^2_{cal} = \frac{\left(a-\frac{a+b}{2}\right)^2}{\frac{a+b}{2}} + \frac{\left(b-\frac{a+b}{2}\right)^2}{\frac{a+b}{2}} = \frac{(a-b)^2}{a+b}$$

On retrouve $X^2_{cal} = (u_{cal})^2$.

VI Régression et corrélation

VI-1 Généralités

Avec le X^2 d'indépendance, nous avons étudié la liaison entre 2 grandeurs *qualitatives* d'un échantillon. Nous allons envisager maintenant le cas des grandeurs *quantitatives*.
On parle alors de *corrélation*. Nous appellons x et y les 2 variables quantitatives.
Différents cas peuvent se présenter :

1 - les variables x et y sont liées l'une à l'autre par une relation ; la connaissance de la valeur x permet de prévoir la valeur y correspondante. On dit qu'il y a une *liaison fonctionnelle* : y = f (x). On obtient graphiquement une courbe en portant les couples de valeurs x_i y_i

2 - les variables x et y n'ont *aucun lien entre elles*.
Si on construit l'ensemble des points, on obtient un nuage de points dispersés tout à fait au hasard entre les axes.

3 - *cas intermédiaire*. Il existe une *dépendance* entre x et y.
Si on construit l'ensemble des points de coordonnées x_i y_i, on obtient un nuage de points qui en gros à la forme d'*une ellipse*. Le nuage est d'autant plus étroit que la liaison statistique est plus serrée (dépendance plus forte).

Il faut considérer 2 cas :

- corrélation *positive ou directe* : x augmente, y augmente en moyenne
- corrélation *négative ou inverse* : x augmente, y diminue en moyenne

VI-2 Droites de régression

Méthode : on réalise un *ajustement linéaire* qui consiste à remplacer le nuage de points par une droite dite droite de régression dans des conditions que nous allons voir.
Soient : n points expérimentaux x_i y_i.
On cherche une droite dont les valeurs y diffèrent peu des y_i.
Pour cela, on utilise la *méthode dite des moindres carrés*. Précisons cela : on cherche la droite d'équation y = a x + b. En un point x_i l'ordonnée de la droite est $y = a x_i + b$.
L'écart avec le point obtenu y_i est donc : $y - y_i = a x_i + b - y_i$
On cherche la droite telle que la somme des carrés des écarts soit minimale. Soit $\Sigma (a x_i + b - y_i)^2$ = minimum.
La droite obtenue est dite droite de régression de y par rapport à x.
On la note : Dy (x).
Nous admettrons que cette droite passe par le *point moyen du nuage* : M_0 ayant pour coordonnées $\bar{x}$ et $\bar{y}$.
Choisissons de nouveaux axes M_0X et M_0Y (parallèles à Ox et à Oy). Dans ce système de coordonnées, la droite de régression a pour équation (simplifiée) : Y = a X
et a est tel que :

$$I = \sum_{i=1}^{n}(aX_i - Y_i)^2 = \text{minimum}$$

Développons :

$$I = \Sigma\,(a^2X_i^2 - 2aX_iY_i + Y_i^2) = a^2\,\Sigma\,X_i^2 - 2\,a\,\Sigma\,X_iY_i + \Sigma\,Y_i^2$$

Cette expression est minimale pour la valeur de a qui annule la dérivée de I par rapport à la variable a ----> $I'_a = 0$

$$I'_a = 2\,a\,\Sigma\,X_i^2 - 2\,\Sigma\,X_iY_i = 0$$

d'où :

$$a = \Sigma\,X_iY_i\,/\,\Sigma\,X_i^2$$

et en revenant à x et y :

$$a = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

On appelle *covariance de x et y* l'expression :

$$\text{cov}\,(x\,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \qquad n = \text{nombre de points}$$

$$\text{mais : var}\,(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

D'où : $a_{y/x} = \dfrac{\text{cov}\,(x,\,y)}{\text{var}(x)}$

($a_{y/x}$ pour préciser qu'il s'agit de droite de régression de y par rapport à x).

Remarque : pour les calculs, on montre facilement que la covariance s'écrit aussi :

$$\text{cov}\,(x, y) = \overline{xy} - \bar{x}\,.\,\bar{y}$$

On détermine de même la *droite de régression de x par rapport à y* en rendant minimale la somme des carrés des distances parallèles à l'axe des x.
On obtient de façon symétrique :

$$a_{x/y} = \frac{\text{cov}\,(x,\,y)}{\text{var}\,(y)}$$

VI-3 Coefficient de corrélation linéaire

On utilise comme paramètre de corrélation, le produit des pentes des 2 droites de régression $a_{y/x} \cdot a_{x/y} = r^2$. r qui est donc la *moyenne géométrique des 2 pentes de régression* est appelé coefficient de corrélation linéaire.
Ceci peut s'écrire en remplaçant $a_{y/x}$ et $a_{x/y}$ par des valeurs vues :

$$\frac{\text{cov}(x, y)}{\text{var}(x)} \quad \frac{\text{cov}(x, y)}{\text{var}(y)} = r^2$$

ou

$$\frac{\text{cov}(x, y)}{s_x^2} \quad \frac{\text{cov}(x, y)}{s_y^2} = r^2$$

soit :

$$r = \frac{\text{cov}(x, y)}{s_x \, s_y}$$

On en déduit les valeurs de $a_{y/x}$ et $a_{x/y}$ correspondantes :

$$a_{y/x} = \frac{\text{cov}(xy)}{s_x^2} = \frac{r \, s_x s_y}{s_x^2} = \frac{r \, s_y}{s_x}$$

de même $a_{x/y} = \frac{r \, s_x}{s_y}$

En définitive, l'équation de la droite de régression de y en x s'écrit :

$$y - \bar{y} = a (x - \bar{x})$$

ou $$y = \bar{y} + a (x - \bar{x})$$

ou $$y = a.\, x + (\bar{y} - a\, \bar{x})$$

avec $$a = a_{y/x} = \frac{\text{cov}(x, y)}{s_x^2} = \frac{r \, s_y}{s_x}$$

On montre que r varie entre -1 et 1.

- quand $|r| = 1$, les 2 droites sont confondues
(relation fonctionnelle = corrélation linéaire parfaite)
- quand r = 0, il n'y a pas corrélation, les 2 droites sont perpendiculaires entre elles et parallèles aux axes.

Remarque importante : de même que pour les autres tests, il faut interpréter avec prudence la corrélation qui existe entre 2 phénomènes. *L'existence d'une corrélation n'implique pas en effet de lien de causalité*. Il peut y avoir une cause commune.

VI-4 Corrélation significativement différente de zéro

Dans bien des cas en biologie, on désire seulement savoir s'il existe ou non une corrélation entre 2 phénomènes étudiés sans être intéressé par la valeur exacte du coefficient de corrélation. On veut donc savoir seulement si la valeur trouvée r diffère significativement de zéro.

a) Test.

- Appelons ρ le coefficient de corrélation de la population (inconnu).
- L'hypothèse H_0 est donc : ρ = 0 (corrélation nulle).
- Le test suppose que la distribution des 2 variables est normale.

Alors sous H_0, on démontre que :

$$\frac{R}{\sqrt{1 - R^2}} \cdot \sqrt{n - 2} = T_{n-2}$$

loi de Student à n - 2 ddl

On calcule donc t_{cal} qu'on compare à $t_{\alpha,n-2}$ correspondant au risque α choisi.

Règle de décision habituelle :

si $|t_{cal}| < t_\alpha$ H_0 acceptée

sinon H_0 rejetée ----> corrélation significativement différente de zéro

b) Emploi de la table

Lorsque l'on cherche si r est significativement différent de 0, on peut utiliser une table spéciale. Cette table construite à partir de la loi de Student fournit la valeur seuil r_α sous l'hypothèse qu'il n'y a pas de corrélation.

r_α dépend du risque α choisi et de n - 2 ddl. Rappelons que l'on suppose la normalité des 2 variables.

VI-5 Intervalle de confiance du coefficient de corrélation

r est une réalisation d'une variable aléatoire R dont les propriétés sont connues *lorsque X et Y suivent une loi normale*.

Soit : ρ la véritable valeur inconnue du coefficient de corrélation de la population.

FISHER a démontré l'intérêt de se ramener à une variable auxiliaire :

$$z = \frac{1}{2} \operatorname{Ln} \frac{1+r}{1-r}$$

z est une réalisation d'une variable Z qui suit approximativement une loi normale d'écart-type σ_Z indépendant de r

$$\sigma_Z = \frac{1}{\sqrt{n-3}}$$

On peut donc connaître les limites de l'intervalle de z donc de r. Il existe une table qui donne la correspondance entre z et r .

VII Analyse de la variance (ANVA)

VII-1 Généralités

Méthode développée par FISHER concernant les séries statistiques doubles. Dans l'étude des séries statistiques doubles, on peut avoir à étudier 3 cas :

- 2 variables quantitatives ---> corrélation
- 2 variables qualitatives ---> X2
- 1 variable qualitative et l'autre quantitative ---> ANVA

Fréquemment en biologie, la valeur qualitative est contrôlée ("facteur" décidé pour l'expérimentation).

L'analyse de la variance permet de tester globalement l'homogénéité des différents échantillons. On peut disposer les résultats sur *k colonnes* (séries) où figurent les valeurs individuelles

VII-2 Le test

Le test va porter sur des sommes de carrés d'écart (SCE) que nous allons préciser.

a) Somme de carrés

Notations nécessaires :

$\bar{x}$	: moyenne générale	$\bar{x}_i$	: moyenne de la série i
x_{ij}	: une valeur individuelle	n_i	: effectif de la série i
T_G	: total général	--->	$\bar{x} = T_G/n$
T_i	: total valeurs série i	--->	$\bar{x}_i = T_i/n_i$

On considère *3 types d'écart* :

1 : $x - \bar{x}$ = écart entre une valeur et la moyenne générale

2 : $\bar{x}_i - \bar{x}$ = écart entre la moyenne de la série et la moyenne générale

3 : $x - \bar{x}_i$ = écart entre une valeur et la moyenne de la série

On définit les *3 sommes correspondantes* où les écarts sont au carré

$$S_T = \sum_{i\,j} (x - \bar{x})^2$$

$$S_f = \sum_{i\,j} (\bar{x}_i - \bar{x})^2 = \sum_{i} n_i (\bar{x}_i - \bar{x})^2 \qquad i = \text{indice de la série}$$

$$S_r = \sum_{i} \sum_{j} (x - \bar{x}_i)^2$$

S_T traduit l'ampleur des variations dans leur ensemble ---> T pour total.

S_f porte uniquement sur les différences entre les séries, c'est-à-dire sur l'influence du facteur étudié ---> f pour facteur.

S_r fait intervenir uniquement les fluctuations à l'intérieur de chaque série ---> r pour résiduel.

b) Relation entre les sommes

On démontre facilement que :

$$S_T = S_R + S_F$$

c) Résultat fondamental

Hypothèse H_0 : les échantillons proviennent d'une même population (= pas d'effet du facteur contrôlé) $\mu_1 = \mu_2 = ...= \mu_k$.

Hypothèses complémentaires : échantillons gaussiens, de même variance (σ^2)

Sous ces hypothèses, on démontre que :

$$f_{cal} = \frac{\frac{S_f}{k-1}}{\frac{S_r}{n-k}}$$

est une réalisation d'une variable de Snedecor F à k - 1 et n - k ddl.

d) Terminologie

$\frac{S_f}{k-1} = s_f^2$ = variance factorielle ou interclasse ou intercolonne

$\frac{S_r}{n-k} = s_r^2$ = variance résiduelle ou intraclasse ou intracolonne

$\frac{S_T}{n-1} = s_T^2$ = variance totale

e) Formules dérivées pour les calculs

Nous avons vu que :

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

On obtient aisément de la même façon les valeurs S_T et S_f.

Pour simplifier $x_{ij} = x$ ---> $S_T = \sum x^2 - n\bar{x}^2$ $\qquad \bar{x} = \frac{T_G}{n}$ ---> $\sum x^2 - \frac{T_G^2}{n}$

$$S_f = \sum n_i \bar{x}_i^2 - n\bar{x}^2 \qquad \bar{x}_i = \frac{T_i}{n_i} \text{ ---> } \sum \frac{T_i^2}{n_i} - \frac{T_G^2}{n}$$

$$S_r = S_T - S_f$$

f) Test

Sous les hypothèses vues, on calcule :

S_T, S_f ----> S_r par différence puis $f_{cal} = \frac{s_f^2}{s_r^2} = \frac{\frac{S_f}{k-1}}{\frac{S_r}{n-k}}$

Au risque α choisi, on cherche dans la table de Snedecor P = α la valeur-seuil f_α qui dépend de α, k - 1 et n -k ddl.

Règle de décision :

si $f_{cal} < f_\alpha$ H_0 acceptée

sinon H_0 rejetée

VIII Les tests de rang

VIII-1 Généralités

Lorsque les effectifs sont petits, le test de comparaison de 2 moyennes de Student-Fisher nécessite *certaines hypothèses* : normalité des distributions, égalité des variances... *Et ceci d'autant plus que les effectifs sont plus faibles.*
Mais malheureusement, c'est dans ces cas là, qu'elles sont très difficiles voire impossibles à vérifier ($n < 10$ et risque que ß soit grand).

C'est pourquoi, on a été conduit à mettre au point toute une série de méthodes qui permettent de traiter les problèmes sans rien exiger des lois de probabilité des variables. Ce sont les méthodes non paramétriques.

Il existe de très nombreux tests non paramétriques ; la plupart sont des *tests de rangs.*
Nous étudierons trois d'entre eux. Ce sont des tests d'emploi simple (de plus en plus utilisés en médecine).

VIII-2 Test U de Mann et Whitney

a) Méthode.

Le test permet de résoudre le problème de comparaison de 2 moyennes d'échantillons indépendants.

Ex : comparer les *observations rangées par ordre croissant*

observations x_1 11, 21, 25, 52, 71, 79 $n_1 = 6$

x_2 22, 43, 72, 91, 116 $n_2 = 5$

total $n = n_1 + n_2 = 6 + 5 = 11$

On range les 11 observations par valeurs croissantes. Soit :

11	21	22	25	43	52	71	72	79	91	116
x_1	x_1	x_2	x_1	x_2	x_1	x_1	x_2	x_1	x_2	x_2

On peut déterminer différents indices. Notamment celui-ci noté U (rien à voir avec la variable normale).
U_1 obtenu ainsi : on compte pour chaque x_1 le nombre des x_2 qui lui sont inférieurs et on fait la somme des résultats.

Ex : $U_1 = 0 + 0 + 1 + 2 + 2 + 3 = 8$

Vous pouvez vérifier que $U_2 = 22$ et que $U_1 + U_2 = n_1 n_2 = 6 . 5 = 30$

Un indice U_1 peut prendre toutes les valeurs comprises entre 0 (tous les x_1 inférieurs à tous les x_2) et n_1n_2 (cas inverse). La valeur moyenne est $\frac{n_1n_2}{2}$

Test : hypothèse H_0 : échantillons extraits d'une même population.

Sous H_0, on montre que (U représente aussi bien U_1 que U_2) :

- loi de U → loi normale : approximation valable quand n_1 et $n_2 > 10$

- E (U) = $\frac{n_1 n_2}{2}$

- Var (U) = $n_1 n_2 (n+1) / 12$

On calcule donc :

$$u_{cal} = \frac{U-E(U)}{\sigma_U} = \frac{U-\frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}}$$

$\left| u_{cal} \right|$ comparé à u_α. Règle habituelle

b) Tables

Elles vont de $n_1 = 1$ à $n_1 = 20$ (attention n_1 = plus petit des effectifs)

et de $n_2 - n_1 = 0$ à 26 ----> (1 à 46 pour n_2)

Elles sont fournies pour une valeur α donnée. On considère le U le plus petit (U min).
Attention : la différence est significative si $U_{min} \leq U_\alpha$ (inhabituel, en général > valeur seuil) (se souvenir que U = 0 ----> séparation complète des 2 échantillons).

Remarques :

1 - le test est un peu moins puissant que le test t quand les distributions sont normales (95 % de puissance).

2 - cas des ex-aequo : on utilise un rang moyen, il ne faut pas qu'ils soient trop nombreux. On compte 1/2 pour tout couple (x, y).

VIII-3 Test T de Wilcoxon

a) Méthode

On utilisera ce test lorsque les séries sont *appariées.*
Comme avec le test t sur valeurs appariées, on forme pour chaque paire la différence.
Particularité : on considère seulement les *différences non nulles.*
On classe les différences par valeur absolue croissante.
Puis on calcule les rangs (rang moyen si ex-aequo).
Soit M = somme des rangs des différences de signe "-";
et P = somme des rangs des différences de signe "+".
On peut vérifier que M + P = n(n+1)/2 (n = nombre de différences ≠ 0)

Hypothèse H_0 : même population. Sous H_0 : on montre que :

- loi de M ----> loi normale : approximation valable quand n > 20.

- E (M) = $\frac{1}{2}\left(\frac{n (n+1)}{2}\right)$

- Var (M) = n (n + 1) (2n + 1)/24

On calcule donc si n > 20 :

$$u_{cal} = \frac{M - E(M)}{\sigma_M} = \frac{M - \frac{n(n+1)}{4}}{\sqrt{\frac{1}{24} n(n+1)(2n+1)}}$$

$\left| u_{cal} \right|$ est comparé à u_α avec la règle de décision habituelle.

b) Table

La table fournie va de n = 0 à n = 20 et pour α = 5 % et α = 1 %.
On considère la plus petite des 2 valeurs M ou P. Soit T cette valeur.
On lit la valeur seuil correspondant à n et α.
Résultat significatif (H_0 rejetée) si T ≤ seuil (inhabituel < ou =).

VIII-4 Coefficient de corrélation des rangs de Spearman

Les données de base sont n couples (x_i, y_i). On classe *séparément* les x et les y.

- à chaque valeur x correspond ainsi un rang de 1 à n.
- de même pour les y.

On substitue à chaque valeur x ou y son rang et on calcule sur ces variables nouvelles (x', y') le coefficient de corrélation habituel.
Simplification : car les x' et y' sont des valeurs entières de 1 à n et la formule habituelle peut être simplifiée. On obtient la formule suivante :

$$r' = 1 - \frac{6 \Sigma d_i^2}{n(n^2 - 1)}$$

où $d_i = y'_i - x'_i$ (= différence de rangs)

Quand n ≥ 10, on teste la significativité de r' comme celle de r (calcul ou table du coefficient de corrélation).
Quand n < 10, on utilise une table spéciale. La table va de n = 5 à n = 10 et α = 5 % et α = 1 % (rien de significativement différent 0 si moins de 5 couples de points).

VIII Les tables

1 - Table de la distribution de POISSON

2 - Loi normale : table de l'écart-réduit

3 - Loi normale : table de la fonction de répartition

4 - Table de la loi du Chi deux

5 - Table de la loi de STUDENT

6 - Table de la loi de SNEDECOR à 2,5%

7 - Table de la loi de SNEDECOR à 5%

8 - Table du coefficient de corrélation

9 - Table de conversion en z du coefficient de corrélation

10 - Table de MANN et WHITNEY

1 - Table de la distribution de POISSON

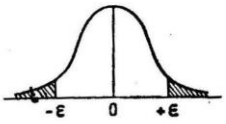
$$\Pr(X = k) = e^{-a}\frac{a^k}{k!}$$

k \ a	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	2,0
0	0,905	0,819	0,741	0,670	0,607	0,549	0,497	0,449	0,407	0,368	0,135
1	0,090	0,164	0,222	0,268	0,303	0,329	0,348	0,359	0,366	0,368	0,271
2	0,005	0,016	0,033	0,054	0,076	0,099	0,122	0,144	0,165	0,184	0,271
3	-	0,001	0,003	0,007	0,013	0,020	0,028	0,038	0,049	0,061	0,180
4	-	-	-	0,001	0,002	0,003	0,005	0,008	0,011	0,015	0,090
5	-	-	-	-	-	-	0,001	0,001	0,002	0,003	0,036
6	-	-	-	-	-	-	-	-	-	0,001	0,012
7	-	-	-	-	-	-	-	-	-	-	0,003
8	-	-	-	-	-	-	-	-	-	-	0,001

k \ a	2,5	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0	14,0	18,0
0	0,082	0,050	0,018	0,007	0,002	0,001	-	-	-	-	-
1	0,205	0,149	0,073	0,034	0,015	0,006	0,003	0,001	-	-	-
2	0,257	0,224	0,147	0,084	0,045	0,022	0,011	0,005	0,002	-	-
3	0,214	0,224	0,195	0,140	0,089	0,052	0,029	0,015	0,008	-	-
4	0,134	0,168	0,195	0,175	0,134	0,091	0,057	0,034	0,019	0,001	-
5	0,067	0,101	0,156	0,175	0,161	0,128	0,092	0,061	0,038	0,004	-
6	0,028	0,050	0,104	0,146	0,161	0,149	0,122	0,091	0,063	0,009	0,001
7	0,010	0,022	0,060	0,104	0,138	0,149	0,140	0,117	0,090	0,017	0,002
8	0,003	0,008	0,030	0,065	0,103	0,130	0,140	0,132	0,113	0,030	0,004
9	0,001	0,003	0,013	0,036	0,069	0,101	0,124	0,132	0,125	0,047	0,008
10	-	0,001	0,005	0,018	0,041	0,071	0,099	0,119	0,125	0,066	0,015
11	-	-	0,002	0,008	0,023	0,045	0,072	0,097	0,114	0,084	0,025
12	-	-	0,001	0,003	0,011	0,026	0,048	0,073	0,095	0,098	0,037
13	-	-	-	0,001	0,005	0,014	0,030	0,050	0,073	0,106	0,051
14	-	-	-	-	0,002	0,007	0,017	0,032	0,052	0,106	0,066
15	-	-	-	-	-	0,003	0,009	0,019	0,035	0,099	0,079
16	-	-	-	-	-	0,001	0,005	0,011	0,022	0,087	0,088
17	-	-	-	-	-	0,001	0,002	0,006	0,013	0,071	0,094
18	-	-	-	-	-	-	0,001	0,003	0,007	0,055	0,094
19	-	-	-	-	-	-	-	0,001	0,004	0,041	0,089
20	-	-	-	-	-	-	-	0,001	0,002	0,029	0,080
21	-	-	-	-	-	-	-	-	0,001	0,019	0,068
22	-	-	-	-	-	-	-	-	-	0,012	0,056
23	-	-	-	-	-	-	-	-	-	0,007	0,044
24	-	-	-	-	-	-	-	-	-	0,004	0,032
25	-	-	-	-	-	-	-	-	-	0,002	0,024
26	-	-	-	-	-	-	-	-	-	0,001	0,016
27	-	-	-	-	-	-	-	-	-	0,001	0,011
28	-	-	-	-	-	-	-	-	-	-	0,007
29	-	-	-	-	-	-	-	-	-	-	0,004
30	-	-	-	-	-	-	-	-	-	-	0,003
31	-	-	-	-	-	-	-	-	-	-	0,002
32	-	-	-	-	-	-	-	-	-	-	0,001
33	-	-	-	-	-	-	-	-	-	-	0,001

2 - Loi normale : table de l'écart-réduit

La table donne la probabilité α pour que l'écart-réduit égale ou dépasse, en valeur absolue, une valeur donnée ε, c'est-à-dire la probabilité extérieure à l'intervalle $(-\varepsilon, +\varepsilon)$.



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,20	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,30	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,40	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,50	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,60	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,70	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,80	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,90	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

La probabilité α s'obtient par addition des nombres inscrits en marge.

Exemple : pour $\varepsilon = 1{,}960$ la probabilité est $\alpha = 0{,}00 + 0{,}05 = 0{,}05$.

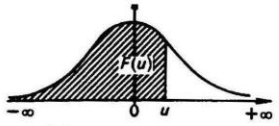
Table pour les petites valeurs de la probabilité.

α	0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
ε	3,29053	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

3 - Loi normale : table de la fonction de répartition

La table donne la probabilité de trouver une valeur inférieure à u.

Lorsque u est négatif, il faut prendre le complément à 1 de la valeur lue.



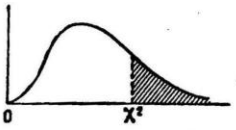
u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,680 8	0,684 4	0,687 9
0,5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 0	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 4	0,773 4	0,776 4	0,779 4	0,782 3	0,785 2
0,8	0,788 1	0,791 0	0,793 9	0,796 7	0,799 5	0,802 3	0,805 1	0,807 8	0,810 6	0,813 3
0,9	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9
1,0	0,841 3	0,843 8	0,846 1	0,848 5	0,850 8	0,853 1	0,855 4	0,857 7	0,859 9	0,862 1
1,1	0,864 3	0,866 5	0,868 6	0,870 8	0,872 9	0,874 9	0,877 0	0,879 0	0,881 0	0,883 0
1,2	0,884 9	0,886 9	0,888 8	0,890 7	0,892 5	0,894 4	0,896 2	0,898 0	0,899 7	0,901 5
1,3	0,903 2	0,904 9	0,906 6	0,908 2	0,909 9	0,911 5	0,913 1	0,914 7	0,916 2	0,917 7
1,4	0,919 2	0,920 7	0,922 2	0,923 6	0,925 1	0,926 5	0,927 9	0,929 2	0,930 6	0,931 9
1,5	0,933 2	0,934 5	0,935 7	0,937 0	0,938 2	0,939 4	0,940 6	0,941 8	0,942 9	0,944 1
1,6	0,945 2	0,946 3	0,947 4	0,948 4	0,949 5	0,950 5	0,951 5	0,952 5	0,953 5	0,954 5
1,7	0,955 4	0,956 4	0,957 3	0,958 2	0,959 1	0,959 9	0,960 8	0,961 6	0,962 5	0,963 3
1,8	0,964 1	0,964 9	0,965 6	0,966 4	0,967 1	0,967 8	0,968 6	0,969 3	0,969 9	0,970 6
1,9	0,971 3	0,971 9	0,972 6	0,973 2	0,973 8	0,974 4	0,975 0	0,975 6	0,976 1	0,976 7
2,0	0,977 2	0,977 9	0,978 3	0,978 8	0,979 3	0,979 8	0,980 3	0,980 8	0,981 2	0,981 7
2,1	0,982 1	0,982 6	0,983 0	0,983 4	0,983 8	0,984 2	0,984 6	0,985 0	0,985 4	0,985 7
2,2	0,986 1	0,986 4	0,986 8	0,987 1	0,987 5	0,987 8	0,988 1	0,988 4	0,988 7	0,989 0
2,3	0,989 3	0,989 6	0,989 8	0,990 1	0,990 4	0,990 6	0,990 9	0,991 1	0,991 3	0,991 6
2,4	0,991 8	0,992 0	0,992 2	0,992 5	0,992 7	0,992 9	0,993 1	0,993 2	0,993 4	0,993 6
2,5	0,993 8	0,994 0	0,994 1	0,994 3	0,994 5	0,994 6	0,994 8	0,994 9	0,995 1	0,995 2
2,6	0,995 3	0,995 5	0,995 6	0,995 7	0,995 9	0,996 0	0,996 1	0,996 2	0,996 3	0,996 4
2,7	0,996 5	0,996 6	0,996 7	0,996 8	0,996 9	0,997 0	0,997 1	0,997 2	0,997 3	0,997 4
2,8	0,997 4	0,997 5	0,997 6	0,997 7	0,997 7	0,997 8	0,997 9	0,997 9	0,998 0	0,998 1
2,9	0,998 1	0,998 2	0,998 2	0,998 3	0,998 4	0,998 4	0,998 5	0,998 5	0,998 6	0,998 6

Table pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
F(u)	0,998 65	0,999 04	0,999 31	0,999 52	0,999 66	0,999 76	0,999 841	0,999 928	0,999 968	0,999 997

4 - Table de la loi du Chi deux

La table donne la probabilité α pour que χ^2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



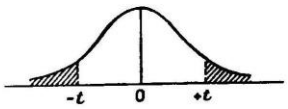
d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,790	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	10,865	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	11,651	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	12,443	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	13,240	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	14,041	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	14,848	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	15,659	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	16,473	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	17,292	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	18,114	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	18,939	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	19,768	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	20,599	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703

Exemple : pour ddl = 1 et $\alpha = 0,05$, on a $X^2 = 3,841$.

Quand ddl > 30, X^2 est distribué à peu près normalement autour de ddl avec une variance égale à 2 ddl.

5 - Table de la loi de STUDENT

La table donne la probabilité α pour que t égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Exemple : pour ddl = 5 et $\alpha = 0,05$, on a t = 2,571.

6 - Table de la loi de SNEDECOR à 2,5%

La table donne la limite supérieure de F pour le risque 2,5 % (valeur ayant 2,5 chances sur 100 d'être égale ou dépassée) en fonction des degrés de liberté k_1 et k_2.

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	∞
1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1008	1013	1016	1017	1018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,3	14,2	14,1	14,0	14,0	13,9	13,9	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,46	8,38	8,32	8,29	8,27	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,23	6,14	6,08	6,05	6,03	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,07	4,98	4,92	4,88	4,86	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,36	4,28	4,21	4,18	4,16	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,89	3,81	3,74	3,70	3,68	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,56	3,47	3,40	3,37	3,35	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,31	3,22	3,15	3,12	3,09	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,33	3,23	3,12	3,03	2,96	2,92	2,90	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,18	3,07	2,96	2,87	2,80	2,76	2,74	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,05	2,95	2,84	2,74	2,67	2,63	2,61	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	2,95	2,84	2,73	2,64	2,56	2,53	2,50	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,64	2,55	2,47	2,44	2,41	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,79	2,68	2,57	2,47	2,40	2,36	2,33	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,72	2,62	2,50	2,41	2,33	2,29	2,26	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,67	2,56	2,44	2,35	2,27	2,23	2,20	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,62	2,51	2,39	2,30	2,22	2,18	2,15	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,35	2,25	2,17	2,13	2,10	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,50	2,39	2,27	2,17	2,09	2,05	2,02	2,00
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,44	2,33	2,21	2,11	2,02	1,98	1,95	1,94
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,39	2,28	2,16	2,05	1,97	1,92	1,90	1,88
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,34	2,23	2,11	2,01	1,92	1,88	1,85	1,83
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,07	1,97	1,88	1,84	1,81	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,18	2,07	1,94	1,83	1,74	1,69	1,66	1,64
50	5,34	3,98	3,39	3,06	2,83	2,67	2,55	2,46	2,38	2,32	2,11	1,99	1,87	1,75	1,66	1,60	1,57	1,55
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,06	1,94	1,82	1,70	1,60	1,54	1,51	1,48
80	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,36	2,28	2,21	2,00	1,88	1,75	1,63	1,53	1,47	1,43	1,40
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	1,97	1,85	1,71	1,59	1,48	1,42	1,38	1,35
200	5,10	3,76	3,18	2,85	2,63	2,47	2,35	2,26	2,18	2,11	1,90	1,78	1,64	1,51	1,39	1,32	1,27	1,23
500	5,05	3,72	3,14	2,81	2,59	2,43	2,31	2,22	2,14	2,07	1,86	1,74	1,60	1,46	1,34	1,25	1,19	1,14
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,83	1,71	1,57	1,43	1,30	1,21	1,13	1,00

Exemple : pour $k_1 = 6$ et $k_2 = 10$, la limite supérieure de F est 4,07.

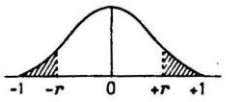
7 - Table de la loi de SNEDECOR à 5%

La table donne la limite supérieure de F pour le risque 5 % (valeur ayant 5 chances sur 100 d'être égale ou dépassée) en fonction des degrés de liberté k_1 et k_2.

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	253	254	254	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,62	8,58	8,55	8,54	8,53	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,70	5,66	5,65	5,64	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,44	4,41	4,39	4,37	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,75	3,71	3,69	3,68	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,38	3,32	3,27	3,25	3,24	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,02	2,97	2,95	2,94	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,80	2,76	2,73	2,72	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,64	2,59	2,56	2,55	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,57	2,51	2,46	2,43	2,42	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,47	2,40	2,35	2,32	2,31	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,38	2,31	2,26	2,23	2,22	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,31	2,24	2,19	2,16	2,14	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,18	2,12	2,10	2,08	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,19	2,12	2,07	2,04	2,02	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,15	2,08	2,02	1,99	1,97	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,11	2,04	1,98	1,95	1,93	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,07	2,00	1,94	1,91	1,89	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,97	1,91	1,88	1,86	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,15	2,07	1,98	1,91	1,85	1,82	1,80	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,11	2,03	1,94	1,86	1,80	1,77	1,75	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,07	1,99	1,90	1,82	1,76	1,73	1,71	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,04	1,96	1,87	1,79	1,73	1,69	1,67	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,84	1,76	1,70	1,66	1,64	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,74	1,66	1,59	1,55	1,53	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,69	1,60	1,52	1,48	1,46	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,56	1,48	1,44	1,41	1,39
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,79	1,70	1,60	1,51	1,43	1,38	1,35	1,32
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,57	1,48	1,39	1,34	1,31	1,28
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,72	1,62	1,52	1,41	1,32	1,26	1,22	1,19
500	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,69	1,59	1,48	1,38	1,28	1,21	1,16	1,11
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,46	1,35	1,24	1,17	1,11	1,00

8 - Table du coefficient de corrélation

La table indique la probabilité α pour que le coefficient de corrélation égale ou dépasse, en valeur absolue, une valeur donnée r, c'est-à-dire la probabilité extérieure à l'intervalle $(-r, +r)$, en fonction du nombre de degrés de liberté (d.d.l.).



d.d.l. \ α	0,10	0,05	0,02	0,01
1	0,9877	0,9969	0,9995	0,9999
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

Exemple : pour ddl = 20 et $\alpha = 0{,}05$, on a r = 0,4227.

9 - Table de conversion en z du coefficient de corrélation

$$z = \frac{1}{2} \ln \frac{1 + r}{1 - r}$$

z	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10
0,0	0,010 0	0,020 0	0,030 0	0,040 0	0,050 0	0,059 9	0,069 9	0,079 8	0,089 8	0,099 7
0,1	0,109 6	0,119 4	0,129 3	0,139 1	0,148 9	0,158 6	0,168 4	0,178 1	0,187 7	0,197 4
0,2	0,207 0	0,216 5	0,226 0	0,235 5	0,244 9	0,254 8	0,263 6	0,272 9	0,282 1	0,291 3
0,3	0,300 4	0,309 5	0,318 5	0,327 5	0,336 4	0,345 2	0,354 0	0,362 7	0,371 4	0,380 0
0,4	0,388 5	0,396 9	0,405 3	0,413 6	0,421 9	0,430 1	0,438 2	0,446 2	0,454 2	0,462 1
0,5	0,469 9	0,477 7	0,485 4	0,493 0	0,500 5	0,508 0	0,515 4	0,522 7	0,529 9	0,537 0
0,6	0,544 1	0,551 1	0,558 0	0,564 9	0,571 7	0,578 4	0,585 0	0,591 5	0,598 0	0,604 4
0,7	0,610 7	0,616 9	0,623 1	0,629 1	0,635 1	0,641 1	0,646 9	0,652 7	0,658 4	0,664 0
0,8	0,669 6	0,675 1	0,680 5	0,685 8	0,691 1	0,696 3	0,701 4	0,706 4	0,711 4	0,716 3
0,9	0,721 1	0,725 9	0,730 6	0,735 2	0,739 8	0,744 3	0,748 7	0,753 1	0,757 4	0,761 6
1,0	0,765 8	0,769 9	0,773 9	0,777 9	0,781 8	0,785 7	0,789 5	0,793 2	0,796 9	0,800 5
1,1	0,804 1	0,807 6	0,811 0	0,814 4	0,817 8	0,821 0	0,824 3	0,827 5	0,830 6	0,833 7
1,2	0,836 7	0,839 7	0,842 6	0,845 5	0,848 3	0,851 1	0,853 8	0,856 5	0,859 1	0,861 7
1,3	0,864 3	0,866 8	0,869 2	0,871 7	0,874 1	0,876 4	0,878 7	0,881 0	0,883 2	0,885 4
1,4	0,887 5	0,889 6	0,891 7	0,893 7	0,895 7	0,897 7	0,899 6	0,901 5	0,903 3	0,905 1
1,5	0,906 9	0,908 7	0,910 4	0,912 1	0,913 8	0,915 4	0,917 0	0,918 6	0,920 1	0,921 7
1,6	0,923 2	0,924 6	0,926 1	0,927 5	0,928 9	0,930 2	0,931 6	0,932 9	0,934 1	0,935 4
1,7	0,936 6	0,937 9	0,939 1	0,940 2	0,941 4	0,942 5	0,943 6	0,944 7	0,945 8	0,946 8
1,8	0,947 83	0,948 84	0,949 83	0,950 80	0,951 75	0,952 68	0,953 59	0,954 49	0,955 37	0,956 24
1,9	0,957 09	0,957 92	0,958 73	0,959 53	0,960 32	0,961 09	0,961 85	0,962 59	0,963 31	0,964 03
2,0	0,964 73	0,965 41	0,966 09	0,966 75	0,967 39	0,968 03	0,968 65	0,969 26	0,969 86	0,970 45
2,1	0,971 03	0,971 59	0,972 15	0,972 69	0,973 23	0,973 75	0,974 26	0,974 77	0,975 26	0,975 74
2,2	0,976 22	0,976 68	0,977 14	0,977 52	0,978 03	0,978 46	0,978 88	0,979 29	0,979 70	0,980 10
2,3	0,980 49	0,980 87	0,981 24	0,981 61	0,981 97	0,982 33	0,982 67	0,983 01	0,983 35	0,983 67
2,4	0,983 99	0,984 31	0,984 62	0,984 92	0,985 22	0,985 51	0,985 79	0,986 07	0,986 35	0,986 61
2,5	0,986 88	0,987 14	0,987 39	0,987 64	0,987 88	0,988 12	0,988 35	0,988 58	0,988 81	0,989 03
2,6	0,989 24	0,989 45	0,989 66	0,989 87	0,990 07	0,990 26	0,990 45	0,990 64	0,990 83	0,991 01
2,7	0,991 18	0,991 36	0,991 53	0,991 70	0,991 85	0,992 02	0,992 18	0,992 33	0,992 48	0,992 63
2,8	0,992 78	0,992 92	0,993 06	0,993 20	0,993 33	0,993 46	0,993 59	0,993 72	0,993 84	0,993 96
2,9	0,994 08	0,994 20	0,994 31	0,994 43	0,994 54	0,994 64	0,994 75	0,994 85	0,994 95	0,995 05

Exemple : pour r = 0,917, on a z = 1,57.

10 - Table de MANN et WHITNEY

Table de U pour $\alpha <= 5\ \%$

La table donne la limite inférieure de U, U étant la plus petite des 2 valeurs ; n_1 et n_2 sont les effectifs des 2 séries, n_1 étant le plus petit. Le symbole - signifie que la différence n'est jamais significative (au seuil 5 %).

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

Exemple : pour $n_1 = 5$ et $n_2 - n_1 = 1$ ($n_2 = 6$), la différence est significative avec un risque $\alpha <= 5\ \%$, dès que $U <= 3$.

11 - Tables de WILCOXON et SPEARMAN

Test T de Wilcoxon pour séries appariées. Table de T

La table donne la limite inférieure de T, T étant le plus petit des 2 totaux M (somme des rangs des différences négatives) ou P (somme des rangs des différences positives) pour les risques $\alpha <= 5\ \%$ et $\alpha <= 1\ \%$, en fonction du nombre n de différences non nulles. Le symbole - signifie que la différence n'est jamais significative (au seuil 5 %). Elle ne l'est jamais pour $n < 6$.

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

Table du coefficient de corrélation r' de Spearman (coefficient de corrélation des rangs)

La table donne la limite inférieure de r' en valeur absolue pour les risques $\alpha <= 5\ \%$ et $\alpha <= 1\ \%$, en fonction du nombre n de couples de valeurs. Le symbole - signifie que la différence n'est jamais significative (au seuil 5 %). Elle ne l'est jamais pour $n < 5$.

n	$\alpha \leqslant 5\ \%$	$\alpha \leqslant 1\ \%$
5	1,00	–
6	0,89	1,00
7	0,79	0,93
8	0,74	0,88
9	0,68	0,83
10	0,65	0,79