Test Statistique, Student, ANOVA et corrélation

Objectifs

- Tests statistiques : principe et utilisation avec le test de Student
- Interprétation
- Vérification de la pertinence du test (autre choix de test)
- Analyse de variance à un facteur
- Acteur
- Corrélation dans le cas d'une hypothèse "effet linéaire"

Principe des tests statistiques

Principes généraux

- Thématique de recherche $\rightarrow$ fondamentale, appliquée, végétale, animale, moléculaire, chimique, intégrée, médicale, ...
- Connaissance du sujet permet d'identifier une question
  - pertinente sur le plan scientifique
  - "testable" d'un point de vue expérimental.
- "Testable d'un point de vue expérimental"
  - une hypothèse peut être formulée (réponse hypothétique à la question posée)
  - une expérience (expérimentation) peut être mise en place
- Une expérimentation consiste à **manipuler un facteur** et à **contrôler les autres**.
- Observation sur une variable de l'effet de la manipulation de ce facteur.

Principe des tests statistiques

Exemple

Après avoir annulé les effets du vent et de la pente en faisant l'expérience en salle et sur sol plat, on peux observer les effets que l'appui sur l'accélérateur a sur la vitesse.

Expérience

- L'expérience donne un résultat
  - **positif** : conforme à l'hypothèse du chercheur
  - **nul** : l'hypothèse ne se confirme pas
  - **surprenant** : non nul, mais allant dans le sens contraire à ce que les connaissances actuelles permettaient de prévoir
- A partir de quand considère-t-on que le résultat est positif, nul ou surprenant ? $\rightarrow$ test statistique
- Les statistiques
  - ne connaissent pas la vérité biologique
  - déterminent des probabilités d'apparition d'un type de résultat "par hasard"

Principe des tests statistiques

Raisonnement statistique

- L'expérimentateur :
  - réalise une expérience
  - Obtient un résultat
- À quel facteur incombe ce résultat ? (retournement du problème)
  $\rightarrow$ Probabilité pour que le résultat soit dû au hasard ?
  - Forte probabilité : le résultat est peut-être dû au hasard
  - Faible probabilité : on attribuera le résultat à l'effet du facteur étudié
- Deux problèmes :
  - Comment exprimer, quantifier, ... le résultat pour pouvoir calculer des probabilités dessus ?
  - comment calculer cette probabilité ?

Exemples étudiés

Le t de Student, l'analyse de variance et la corrélation

Le test t de Student pour échantillons indépendant

Étude de l'effet d'un facteur à deux modalités sur une variable dépendante

Utilisation

- Populations étudiées distribuées de façon normale
- variabilités des groupes similaires

Correction de Welsch

- Si les variabilités ne sont pas équivalentes
- La plupart du temps automatique
- Diminue le degré de liberté en fonction de la différence entre les variances

Le test t de Student pour échantillons indépendant

Formule

Quantification de notre façon intuitive de dire que les moyennes des deux groupes sont effectivement différentes. D'autant plus important lorsque :

- La différence entre les deux moyennes est importante
- Les variabilités des données sont faibles
- Le nombre d'individus est élevé

$$t = \frac{|m_1 - m_2|}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Le test t de Student pour échantillons indépendant

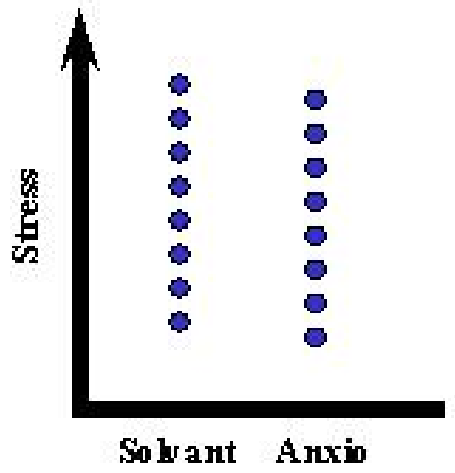
p-value

La *p-value* associée exprime la probabilité pour obtenir par hasard le résultat observé si le facteur n'a pas d'effet (ou si les deux échantillons sont issus de la même population)

- Si $p < 0.05$ on considère que le résultat n'est pas le fruit du hasard : le résultat est significatif
- Sinon, le résultat a, en l'absence d'effet du facteur une telle probabilité d'apparition qu'on ne l'attribuera pas à l'effet du facteur : le résultat n'est pas significatif

Le test t de Student pour échantillons appariés

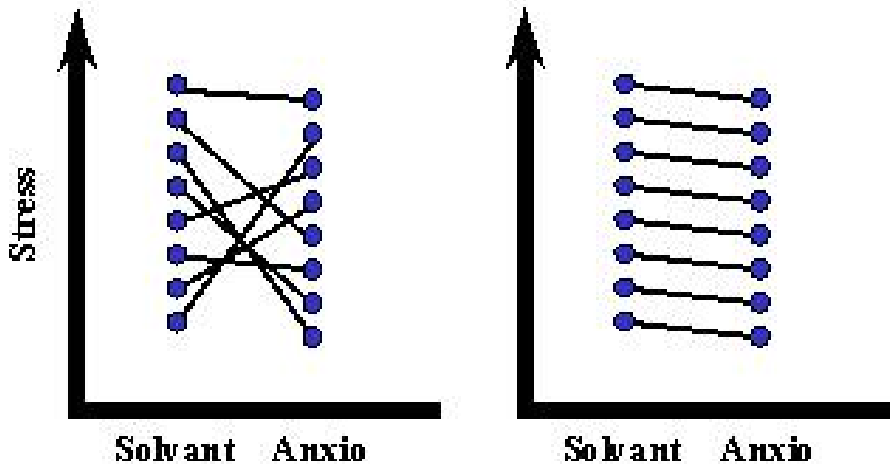
Résultats individuels pour deux groupes, l'un ayant reçu des anxiolytiques et l'autre non. Mesure du stress.



Différence due à l'anxiolytique ou au hasard ?

Le test t de Student pour échantillons appariés

Résultats individuels pour un groupe de sujets dont on évalue le stress avant et après la prise d'anxiolytiques. Sur le premier graphique, la différence est-elle due à



Différence due à l'anxiolytique ou au hasard ?

Il est beaucoup plus facile de répondre dans ces deux cas car on connaît l'évolution de chaque individu.

Le test t de Student pour échantillons appariés

Formule

Calcule des différences entre les deux modalités du facteur pour chaque sujet, puis on compare la moyenne de ces différences avec la valeur "0" (si le facteur n'a pas d'effet, la moyenne des différences va tendre vers 0)

$$t = \frac{|m_d|}{\frac{s_d}{\sqrt{n}}}$$

Exercice 1 : Student

Vos connaissances des systèmes neurobiologiques responsables de la mémoire vous permettent de faire l'hypothèse que la substance X améliore la mémoire. Vous prenez deux groupes de 15 sujets. À un groupe, vous administrez des pastilles contenant la substance X (groupe "Traité"). À l'autre groupe, vous administrez les mêmes pastilles sans la substance X (groupe "Placebo"). Vous faîtes passer un test de mémoire à ces sujets et receuillez les résultats. Votre hypothèse est-elle vérifiée ?

1. Introduisez les données dans R
2. Faites un t de Student sur ces données à l'aide de R
3. Est-ce que R et Systat renvoient les mêmes résultats ?
4. Rédigez le résultat et vos conclusions comme dans un rapport
5. Que vous inspire le graphique ci-dessous ?
6. Quelles vérifications proposez-vous de faire ?

Exercice 1 : Student

SYSTAT :

****

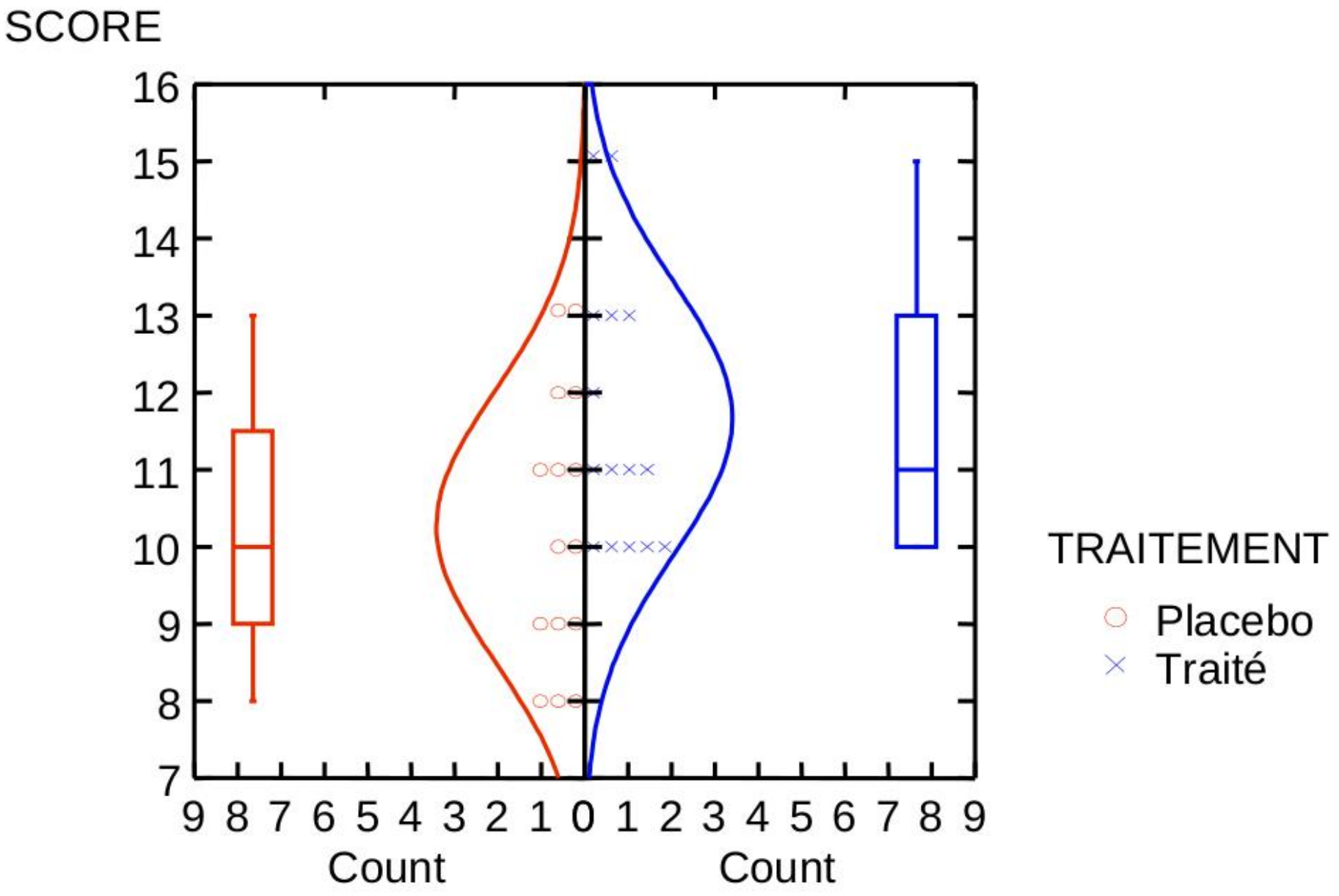
Two-sample t test on SCORE grouped by TRAITEMENT$

Group	N	Mean	SD
Placebo	15	10.267	1.751
Traité	15	11.667	1.759

Separate Variance t = -2.184 df = 28.0 Prob = 0.037
Difference in Means = -1.400 95.00% CI = -2.713 to -0.087

Pooled Variance t = -2.184 df = 28 Prob = 0.037
Difference in Means = -1.400 95.00% CI = -2.713 to -0.087

Exercice 1 : Student



Exercice 2 : Student

On souhaite savoir si un joueur est meilleur que l'autre. Pour chaque joueur, on regarde ses performances aux différents matchs.

1. Qu'en pensez vous en regardant leur moyenne de points ?
   On effectue un test statistique.
2. Introduisez les données dans R
3. Faites un t de Student sur ces données à l'aide de R
4. Est-ce que R et Systat renvoient les mêmes résultats ?
5. Interprétez les résultats et rédigez vos conclusions
6. Vous devez en sélectionner un pour représenter l'équipe : Lequel choisissez-vous ? Pourquoi ?
7. Qui ou que sont les "sujets" (ou "individus") dans cette expérience ?

Exercice 2 : Student

SYSTAT :

****

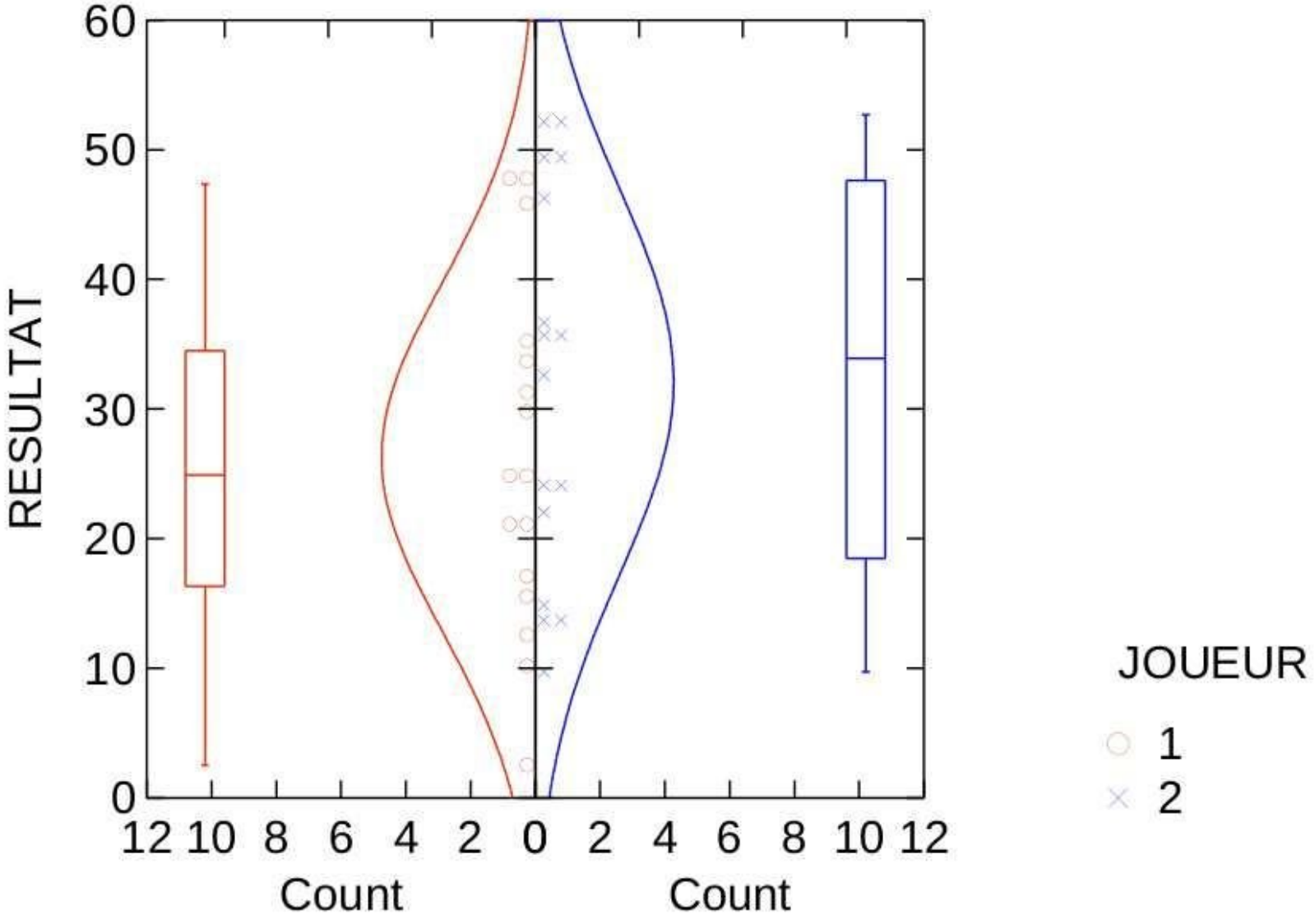
Two-sample t test on RESULTAT grouped by JOUEUR$

Group	N	Mean	SD
1	16	26.296	13.441
2	16	32.107	14.954

Separate Variance t = -1.156 df = 29.7 Prob = 0.257
Difference in Means = -5.810 95.00% CI = -16.081 to 4.460

Pooled Variance t = -1.156 df = 30 Prob = 0.257
Difference in Means = -5.810 95.00% CI = -16.076 to 4.455

Exercice 2 : Student



Analyse de variance pour échantillons indépendants

Utilité

- Dans de nombreuses expériences, il n'y a pas "2" mais "plus de 2" groupes à comparer (Exemple : "l'effet dose" en pharmacologie)
- Effectuer toutes les comparaisons voulues avec des t de Student : chaque t de Student mesure la probabilité d'obtenir par hasard le résultat observé. $\rightarrow$ Multiplie les risque d'en trouver une significative alors que cette significativité est due au hasard (aléas de l'échantillonnage dans notre cas)
- L'analyse de variance permet de corriger ce biais.
- L'analyse de variance est une extension du t de Student lorsqu'on a plus de deux moyennes à comparer : Etude de l'effet d'un facteur à plusieurs modalités sur une variable dépendante $\rightarrow$ Populations étudiées distribuées de façon normale et homoscédasticité

Analyse de variance pour échantillons indépendants

Principes

Analyse en une ou deux étapes suivant les résultats

- Vérifier que globalement la dispersion (des moyennes) des groupes a peu de chance d'être due au hasard. Si c'est le cas :
- Dans ce cas quels sont les groupes qui s'écartent le plus des autres : quelles sont les différences entre les groupes qui ont peu de chances d'être dues au hasard ? $\rightarrow$ Comparaisons "a posteriori", ou "post-hoc" avec le test de Tukey (2ème étape que si la 1ère a montré un effet significatif du facteur).

Autres tests

- Tukey : comparer toutes les moyennes entre elles
- Neuman-Keuls : comparer les moyennes si on a une hypothèse précise sur l'ordre des moyennes
- Dunnett : comparer toutes les moyennes de groupes expérimentaux à un (ou deux) groupes témoins

Analyse de variance en mesure répétées

Principe

De même que pour les t de Student pour échantillons appariés, il existe une procédure plus puissante en ANOVAs pour traiter les répétitions de mesures sur les mêmes sujets. $\rightarrow$ Analyse de variance en mesures répétées (repeated measure ANOVA).

Exercice 3 : Analyse de variance

Vos connaissances au sujet de la structure de l'enzyme W vous permettent de supposer que la molécule Z en est un activateur. Vous disposez de cette molécule Z et vous êtes capable de mesurer l'activité de l'enzyme W. Vous ne savez cependant pas à quelle dose administrer la molécule Z. Vous prenez 5 groupes de 10 souris. Aux différents groupes, vous administrez respectivement 1ng, 10 ng, 50 ng et 100 ng de Z. Le 5e groupe ne reçoit que le solvant utilisé.

1. Introduisez les données dans R
2. Faites une analyse de variance sur ces données à l'aide de R
3. Les résultats sont-ils conformes à ceux de Systat ?
4. Votre hypothèse est-elle vérifiée ?
5. Lisez vous le résultat des tests post-hoc (tests a posteriori) ? Pourquoi ?
6. Rédigez vos conclusions.

Exercice 3 : Analyse de variance

```
SYSTAT :
*******************************************************************************
Categorical values encountered during processing are:
DOSE$ (5 levels)
   000 ng, 001 ng, 005 ng, 010 ng, 100 ng

Dep Var: ACT_ENZ   N: 50   Multiple R: 0.838
Squared multiple R: 0.702
```

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
DOSE$	1.03821E+07	4	2595533.920	26.508	0.000
Error	4406199.300	45	97915.540		

Least squares means.

		LS Mean	SE	N
DOSE$	=000 ng	50748.400	98.952	10
DOSE$	=001 ng	50787.200	98.952	10
DOSE$	=005 ng	50951.300	98.952	10
DOSE$	=010 ng	51332.800	98.952	10
DOSE$	=100 ng	51970.200	98.952	10

Exercice 3 : Analyse de variance

```
  ROW DOSE$
  1  000 ng 2  001 ng 3  005 ng 4  010 ng 5  100 ng
Post Hoc test of ACT_ENZ
-----------------------------------------------------------------------
Using model MSE of 97915.540 with 45 df.
Matrix of pairwise mean differences:
                    1           2           3           4           5
        1           0.000
        2          38.800       0.000
        3         202.900     164.100       0.000
        4         584.400     545.600     381.500       0.000
        5        1221.800    1183.000    1018.900     637.400       0.000

Tukey HSD Multiple Comparisons.
Matrix of pairwise comparison probabilities:
                    1           2           3           4           5
        1           1.000
        2           0.999       1.000
        3           0.599       0.767       1.000
        4           0.001       0.003       0.066       1.000
        5           0.000       0.000       0.000       0.000       1.000
-----------------------------------------------------------------------
```

Exercice 4 : Comparaison ANOVA vs Student

Les données de l'exercice 1 sont traitées avec une ANOVA au lieu du Student

1. Comparez la Prob du test de Student avec le p de l'ANOVA
2. Comparez le t de Student avec le F de l'ANOVA, quelle est la relation qui les unis ?
3. Que concluez-vous ?

Exercice 4 : Comparaison ANOVA vs Student

```
SYSTAT :
*******************************************************************************
Categorical values encountered during processing are:
TRAITEMENT$ (2 levels)
   Placebo, Traité

Dep Var: SCORE   N: 30   Multiple R: 0.382   Squared multiple R: 0.146

                          Analysis of Variance
Source              Sum-of-Squares   df  Mean-Square     F-ratio       P
TRAITEMENT$               14.700     1       14.700       4.771       0.037
Error                     86.267    28        3.081

------------------------------------------------------------------------------
Least squares means.
                                    LS Mean          SE      N
  TRAITEMENT$ =Placebo                10.267       0.453      15
  TRAITEMENT$ =Traité                 11.667       0.453      15
------------------------------------------------------------------------------
```

Le coefficient de corrélation de Pearson

Principe

- Le coefficient de corrélation de Pearson (et son test) est utilisé pour mesurer une relation linéaire entre deux variables quantitative.
- On l'utilise théoriquement lorsque la population étudiée est distribuée de façon normale sur les deux variables.
- Le coefficient de corrélation de Pearson (également appelé coefficient de corrélation de Bravais-Pearson), noté "r", peu prendre les valeurs comprises entre -1 et $+1$.
  - $r = 1$: relation linéaire parfaite, droite de pente positive
  - $r = -1$: relation linéaire parfaite, droite de pente négative
  - $r \simeq 0$: absence de relation linéaire mais il peut y avoir une relation d'un autre type.
  - $-1 < r < 0$: relation linéaire négative : le nuage de points présente une pente descendante.
  - $0 < r < +1$: relation linéaire positive : le nuage de points présente une pente ascendante.

Le coefficient de corrélation de Pearson

Utilisation

Test du coefficient de corrélation : connaître avec quelle probabilité, deux variables qui ne sont pas liées donneront un coefficient tel que celui observé.

- Résultat significatif : les deux variables présentent une relation qui, au minimum, présente une "ressemblance" avec la linéarité.
- Résultat non significatif : le test a été incapable de déceler une relation de linéarité.
  - Pas de relation
  - Indépendance entre les deux variables
  - Relation à composante linéaire mais dont l'intensité est trop faible (masquée par la variabilité induite par les facteurs aléatoires)
  - Relation entre les 2 variables dont la forme empêche la détection d'une relation linéaire (Exemple : relation motivation - performance)

Exercice 5 : Corrélation de Pearson

En psychologie expérimentale, on étudie la relation entre la rapidité dans une tâche et la précision de la réponse. On calcule un coefficient de corrélation de Pearson entre ces deux paramètres.

1. Introduisez les données dans R
2. Calculez le coefficient de corrélation sur ces données à l'aide de R ; ce coefficient est il significatif ?
3. Les résultats sont-ils conformes à ceux de Systat ?
4. Rédigez les conclusions de l'analyse.

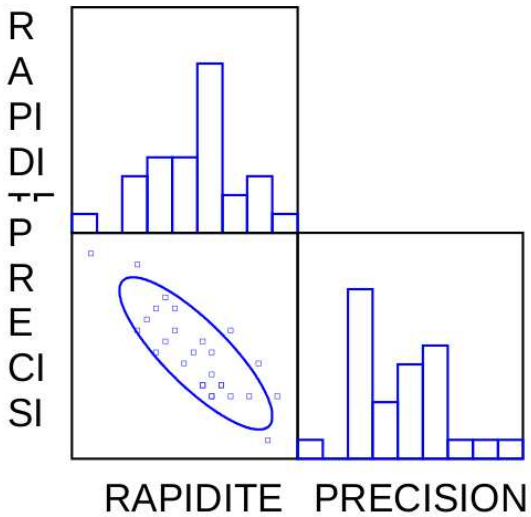
Exercice 5 : Corrélation de Pearson

SYSTAT :

****

Pearson correlation matrix

	RAPIDITE	PRECISION
RAPIDITE	1.000	
PRECISION	-0.824	1.000



Matrix of Probabilities

	RAPIDITE	PRECISION
RAPIDITE	0.000	
PRECISION	0.000	0.000

Number of observations: 27

Exercice 6 : Corrélation de Pearson

Vous savez que l'hippocampe (structure cérébrale) envoie des axones vers le septum latéral. On peut donc s'attendre à observer une relation fonctionnelle. On mesure l'activité de l'hippocampe et celle du septum latéral sur 20 sujets. On va utiliser un coefficient de corrélation pour mesurer cette relation et un test du coefficient de corrélation pour voir si ce résultat peut être attribué aux projections anatomiques.

1. Introduisez les données dans R
2. Calculez le coefficient de corrélation sur ces données à l'aide de R ; ce coefficient est il significatif ?
3. Les résultats de R de Systat sont-ils conformes ?
4. Rédigez les conclusions de l'analyse.

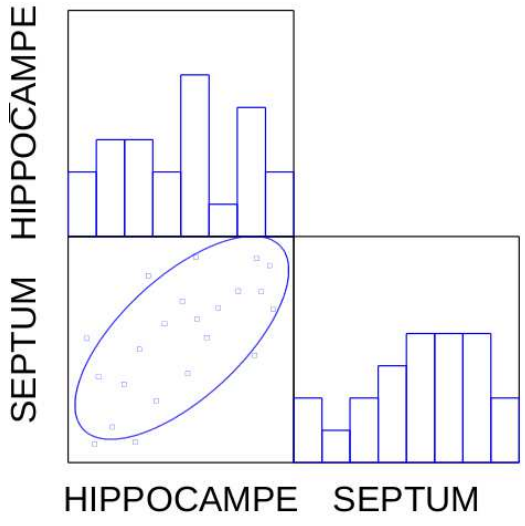
Exercice 6 : Corrélation de Pearson

SYSTAT :

****

Pearson correlation matrix

	HIPPOCAMPE	SEPTUM
HIPPOCAMPE	1.000	
SEPTUM	0.665	1.000



Matrix of Probabilities

	HIPPOCAMPE	SEPTUM
HIPPOCAMPE	0.000	
SEPTUM	0.001	0.000

Number of observations: 22