Fiche TD 2

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 1 facteur, corrélation dans le cas d'une hypothèse " effet linéaire "**

Principe des tests statistiques

[faire trouver un (ou des) exemple(s) concret(s) aux étudiants et le(s) suivre]

Le scientifique a une thématique de recherche (fondamentale, appliquée, végétale, animale, moléculaire, chimique, intégrée, médicale, environnementale...).

Dans cette thématique, sa connaissance du sujet lui permet d'identifier une question qui est à la fois pertinente sur le plan scientifique, et "testable" d'un point de vue expérimental.

"Testable d'un point de vue expérimental" signifie que :

- une hypothèse peut être formulée, c'est-à-dire une réponse hypothétique à la question posée ;
- une expérience (expérimentation) peut être mise en place pour voir si cette hypothèse se vérifie ou pas.

Une expérimentation consiste à manipuler un facteur (la dose d'un traitement,...) et à contrôler les autres (les empêcher d'avoir une influence). Cette démarche permet d'observer sur une variable (un résultat, en général quantitatif), l'effet de la manipulation de ce facteur.

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

L'expérience va donner un résultat qui sera :

- soit positif : conforme à l'hypothèse du chercheur

- soit nul : l'hypothèses ne se confirme pas

- soit surprenant : non nul, mais allant dans le sens contraire à ce que les connaissances actuelles permettaient de prévoir (cas des hypothèses fortes, traitées avec un test unilatéral).

A partir de quand considère-t-on que le résultat est positif, nul ou surprenant ?

C'est ici qu'intervient le test statistique

Les statistiques ne connaissent pas la vérité biologique, par contre elles sont capables de déterminer des probabilités d'apparition d'un type de résultat "par hasard".

Le raisonnement est le suivant :

Ce que fait l'expérimentateur :

- J'ai fait une expérience, j'obtiens un résultat

- ce résultat est-il dû au facteur que j'étudie ?

Pour répondre à cette question, le statisticien retourne le problème :

- Quelle est la probabilité pour que le résultat que j'observe soit dû au hasard ? (c'est-à-dire si le facteur étudié n'a pas d'effet)

- Si cette probabilité est forte, on dira que le résultat est peut-être dû au hasard; si elle est faible, on dira que le hasard a peu de chance de donner ce résultat et on attribuera le résultat à l'effet du facteur étudié.

Nous sommes alors confrontés à deux problèmes :

- Comment exprimer, quantifier... le résultat pour pouvoir calculer des probabilités dessus ?

- Comment calculer cette probabilité ?

On va prendre 3 exemples : le test de Student, l'analyse de variance et la corrélation

Le test t de Student pour échantillons indépendants

"**Le test t de Student est utilisé quand on veut étudier l'effet d'un facteur à deux modalités sur une variable dépendante**" ou en termes plus simples, lorsque l'on veut comparer les moyennes de deux groupes (illustrer par des exemples concrets)

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

On l'utilise lorsque les populations étudiées sont distribuées de façon normale et que les variabilités des groupes sont similaires (dans ce cas on dit qu'il y a homoscédasticité).

Si les variabilités des deux groupes ne sont pas équivalentes, on utilise une correction du test : la correction de Welsch. En général elle est automatiquement effectuée par les bons logiciels. Cette correction consiste à diminuer le degré de liberté de façon plus ou moins importante en fonction de la différence entre les variances.

Le t de Student est une "statistique" qui reflète quantitativement notre façon intuitive de dire que les moyennes des deux groupes sont effectivement différentes. Cette décision est facilitée quand :

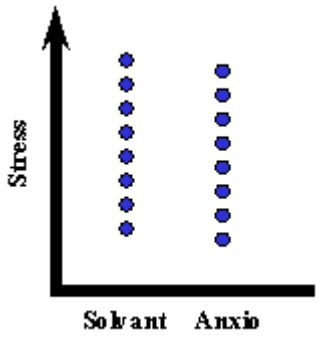
- la différence entre les deux moyennes est importante
- les variabilités des données sont faibles
- le nombre d'individus est élevé

Ce que reflète parfaitement la formule. La valeur du t est corrélée avec notre facilité à dire que la différence est due au facteur étudié.

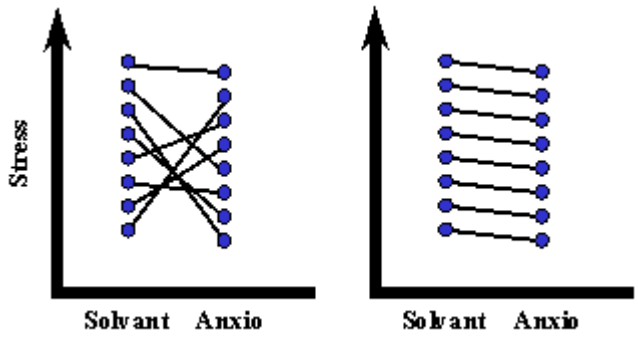
La p-value associée au t exprime la probabilité pour obtenir par hasard le résultat observé si le facteur n'a pas d'effet (ou si les deux échantillons sont issus de la même population). Si cette probabilité est faible (< 0.05) on considère que le résultat n'est pas le fruit du hasard : le résultat est 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

Ci-dessous, les résultats individuels pour un groupe ayant reçu des anxiolytiques et un groupe qui n'en a pas reçu ; on mesure le stress. La différence est-elle due à l'anxiolytique ou au hasard ? Il est difficile de répondre.



Ci-dessous, les 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 l'anxiolytique ou au hasard ? Même question pour le deuxième graphique. Il est beaucoup plus facile de répondre dans ces deux cas car on connaît l'évolution de chaque individu.



Le principe du calcul est à peu près identique au t pour échantillons indépendants sauf que l'on calcule les différence entre les deux modalité 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).

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

L'analyse de variance pour échantillons indépendants

Utilité : dans de nombreuses expériences, il n'y a pas "2" mais "plus de 2" groupes à comparer (exemples de "l'effet dose" en pharmacologie, de plusieurs méthodes à comparer...). La première idée qui nous vient à l'esprit pour résoudre le problème est d'effectuer toutes les comparaisons voulues avec des t de Student. On sait que chaque t de Student mesure la probabilité d'obtenir par hasard (c'est-à-dire si le facteur étudié n'a pas d'effet), le résultat que l'on a observé. Or, utiliser un t de Student sur de nombreuses comparaisons, multiplie les risque d'en trouver une qui est significative alors que cette significativité est due au hasard (aux aléas de l'échantillonnage dans notre cas).
L'analyse de variance permet de corriger ce biais.

Ainsi :
L'analyse de variance est une extension du t de Student lorsqu'on a plus de deux moyennes à comparer.
Elle est utilisée (entre autre) quand on veut étudier l'effet d'un facteur à plusieurs modalités sur une variable dépendante (illustrer par un exemple concret)
On l'utilise lorsque les populations étudiées sont distribuées de façon normale et qu'il y a homoscédasticité.

Dans la pratique on effectuera l'analyse en une ou deux étapes suivant les résultats.
Première étape : vérifier que globalement la dispersion (des moyennes) des groupes a peu de chance d'être due au hasard. Si c'est le cas :
Deuxième étape : on cherche 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 ? On appelle ça des comparaisons "a posteriori", ou "post-hoc". On utilisera pour effectuer ces comparaisons, le test de Tukey (cette deuxième étape n'a lieu que si la première étape a montré un effet significatif du facteur).
Certains logiciels proposent d'autres tests :
- le Tukey est utilisé pour comparer toutes les moyennes entre elles
- le Neuman-Keuls est utilisé pour comparer les moyennes si on a une hypothèse précise sur l'ordre des moyennes
- le Dunnett est utilisé pour comparer toutes les moyennes de groupes expérimentaux à un (ou deux) groupes témoins
- le Scheffé est parfois proposé mais est très peu puissant. Il trouve d'autres utilités intéressantes

L'analyse de variance en mesures répétées

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. Elle s'appelle l'analyse de variance en mesures répétées (repeated measure ANOVA).

Le coefficient de corrélation de Pearson

"**Le coefficient de corrélation de Pearson (et son test) est utilisé pour mesurer une relation linéaire entre deux variables quantitative.**" (phrase à expliquer et à illustrer par exemples concrets)
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 indique une relation linéaire parfaite : tous les points sont situés sur une droite de pente positive
- r = -1 indique une relation linéaire parfaite : tous les points sont situés sur une droite de **pente négative**
- r ≈ 0 indique une absence de relation linéaire mais il peut y avoir une relation d'un autre type.
- -1 < r < 0 indique une relation linéaire négative : le nuage de points présente une pente descendante.
- 0 < r < +1 indique une relation linéaire positive : le nuage de points présente une pente ascendante.

Le test du coefficient de corrélation est utilisé pour connaître avec quelle probabilité, deux variables qui ne sont pas liées donneront un coefficient tel que celui que l'on observe.
Un résultat significatif au test indiquera que les deux variables présentent une relation qui, au minimum, présente une "ressemblance" avec la linéarité.
Un résultat non significatif au test indiquera que le test a été incapable de déceler une relation de linéarité. Soit il n'existe pas de relation entre les deux variables (indépendance entre les variables), soit les deux variables présentent une relation à composante linéaire mais dont l'intensité est trop faible. Elle sera donc masquée par la variabilité induite par les facteurs aléatoires. Soit il existe une relation entre les deux variables mais la forme de la relation empêche la détection d'une relation linéaire (exemple de la relation motivation - performance).