Modèles statistiques

Frédéric Bertrand[1]

[1] IRMA, Université Louis Pasteur
Strasbourg, France

Master 1ère Année 27-09-2006

Au cours de cet enseignement, il vous faudra toujours garder à l'esprit que ce que vous allez apprendre à faire vous serez **tous sans exception** amenés à l'utiliser au cours des deux stages que compte votre formation.

En effet, votre discipline repose sur l'analyse de résultats expérimentaux qui sont donc sujets à des variations aléatoires que vous ne pouvez pas maîtriser. Vous allez néanmoins essayer d'utiliser ces résultats pour formuler puis confirmer ou infirmer des hypothèses.

Les outils statistiques que nous allons voir et utiliser ensemble vous permettront() d'essayer d'extraire le bruit, c'est-à-dire les pertubations indésirables, des informations que vous recueillerez.

Par exemple l'an dernier **tous** les étudiants de deuxième année ont réalisé une analyse statistique des études comportementales qu'ils avaient réalisées au cours de leur stage final.

Tous n'ont pas rencontré les mêmes difficultés durant la phase de traitement de leurs données. Ces différences ne s'expliquent pas toutes par la diversité des sujets traités. En effet la manière dont certains ont récolté leurs données ne leur permettait de se servir d'aucune technique de statistique qu'ils avaient apprise. Il faudra donc qu'au moment **de concevoir** votre expérience vous vous soyez déjà renseignés sur la manière dont vous allez « faire parler »les résultats expérimentaux.

Sommaire

Ce premier cours a pour but de faire un rapide exposé de ce que l'on appelle un modèle statistique tout en vous permettant de revoir, ou de découvrir, certaines des notions que vous avez apprises au cours des trois années de la licence.

Nous allons revoir ensemble :

- L'analyse de la variance à 1 facteur
- Le test non-paramétrique de Krukal-Wallis
- La régression linéaire simple

Sommaire

Relations

Pourquoi a-t-on besoin des statistiques pour analyser des résultats expérimentaux ?

Il existe plusieurs types de relations en des grandeurs physiques comme la masse, la taille, la température...

On en distingue principalement deux :

- les relations **déterministes** comme celle qui lie l'expression d'une température en dégré Celsius et l'expression de cette même température en Kelvin. Ici rien de plus mystérieux qu'une addition à faire et étant donné une même température de départ le résultat sera toujours le même.

- les relations **stochastiques** comme celle qui lie la masse d'un idividu à sa taille. On ne peut pourtant pas nier qu'il y a une association entre la taille et la masse d'une personne mais celle-ci n'est pas aussi simple que celle ci-dessus. En effet si vous comparez la masse de deux personnes qui ont la même taille il est fort probable que celles-ci diffèrent.

Pourtant une telle relation existe. Comment peut-on alors la mettre en évidence ?

Pour mettre en équation la relation du transparent précédent entre le poids et la masse on écrit :

$$Masse(Individu) = Fonction(Taille_{Individu}) + Erreur(Individu)$$

Ce que l'on appelle *Erreur* représente la variabilité inter-individu, c'est-à-dire ce qui permet d'expliquer pourquoi deux personnes de même taille n'auront pas la même masse.

Problème

Comment trouver *Fonction* et *Erreur* ?

On ne connaîtrait vraiment *Fonction* et *Erreur* que si l'on réalisait une infinité d'expérience !

C'est pourquoi en statistique on adopte la démarche opposée :

Réponse

On va **proposer** des canditats pour *Fonction* et *Erreur* puis évaluer l'**adéquation** du modèle proposé avec la réalité.

Le cas de l'erreur

Jusqu'à présent on vous a sans doute dit d'utiliser des erreurs qui suivent des lois normales, mais savez-vous pourquoi ?

Dans beaucoup de problèmes expérimentaux l'erreur qui vient perturber le résultat d'une expérience est la somme de plus petites erreurs du même ordre et indépendantes.

Un théorème de probabilité, le **théorème central limite**, que vous avez dû rencontrer en L2, nous dit qu'alors une bonne approximation de la loi de la variable aléatoire d'**erreur** peut être réalisée en utilisant une **loi normale**.

Bien entendu ceci ne fonctionne pas à tous les coups et il existe alors des alternatives :

- changer la loi de l'erreur
- utiliser des tests comme celui de Kruskal-Wallis qui ne fait quasiment pas d'hypothèse sur la loi des erreurs.

Quelle fonction utiliser ?

Quelles fonctions peut-on utiliser ?
Appelons Y la réponse observée, ou facteur expliqué, et X le facteur explicatif.

$$Y = f(X) + \epsilon$$

La réponse à cette question dépend avant tout de la nature de la variable X.

- Si X est une variable continue comme le poids ou la taille on pourra utiliser $f(X) = a * X + b$.
- Si X est une variable discrète on utilisera plutôt, en notant X_i les différentes valeurs possibles pour X, $f(X_i) = \mu + \alpha_i$.

Synthèse

On résume ce que l'on vient de voir à propos des modèles statistiques.

- Si X est continue on s'intéressera à une relation du type

$$Y = a * X + b + \epsilon$$

  où les variables d'erreurs ϵ suivent toute une loi normale. Cette situation est celle de l'analyse de régression simple.

- Si X est discrète on s'intéressera à une relation du type

$$Y = \mu + X_i + \epsilon$$

  où les variables d'erreurs ϵ suivent toute une loi normale. Cette situation est celle de l'ANOVA à un facteur contrôlé.

Sommaire

Introduction

Supposons que l'on mesure plusieurs fois une même grandeur on trouve en général des résultats différents. De très nombreux facteurs peuvent influencer les résultats et il n'est pas possible de tous les étudier. On en sélectionne un certain nombre : on retiendra ainsi ceux qui a priori peuvent justifier une grande part de la dispersion des mesures.

Ces facteurs sur lesquels nous fixons notre attention seront dits facteurs contrôlés. Ceci implique qu'avant d'effectuer les mesures on aura pris des dispositions pour qu'ils soient maintenus constants et mesurés.

Pour l'instant nous ne nous intéressons qu'au cas où il y a un seul facteur contrôlé.

L'expérimentateur peut se poser alors différentes questions :

- Le phénomène étudié est-il ou non inluencé par le facteur contrôlé ?
- Si la réponse est affirmative, quelle est alors la modalité la plus intéressante ?

Modèle statistique

Le modèle s'écrit, en notant $y_{i,j}$ la $j$ème mesure obtenue au $i$ème niveau du facteur X :

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j}, \quad i = 1, \ldots I, \; j = 1, \ldots, J.$$

avec les hypothèses suivantes pour les résidus :

$$\forall\ (i,j)\ \mathcal{L}(\epsilon_{i,j}) = \mathcal{N}(0, \sigma^2) \text{ et } Cov(\epsilon_{i,j}, \epsilon_{r,s}) = 0 \text{ si } (i,j) \neq (r,s)$$

Notez qu'ici le plan de l'ANOVA est dit **équilibré** car il y a le même nombre de répétitions pour tous les niveaux du facteur. Vous verrez en exercice, un cas de plan déséquilibré.

Dans le slide précédent, il est à noter que $y_{i,j}$ n'est pas égal à $Y_{i,j}$. Dans le premier cas $y_{i,j}$ est une valeur mesurée dans le second cas $Y_{i,j}$ est la variable aléatoire.

Cette remarque a déjà été faite dans les années passées mais il vaut mieux le mentionner ici.

On voit ainsi que l'on fait plusieurs hypothèses très **importantes**. Dans chaque cas où vous essayerez d'utiliser cette outil statistique vous **DEVREZ** vérifier que les hypothèses que l'on fait sont compatibles avec les données expérimentales dont vous disposez.

Si vous utilisez un outil statistique alors qu'il n'est pas adapté vous obtiendrez des résultats qui peuvent être trompeurs voire complètement **faux**.

Votre logiciel de calcul statistique ne s'occupera pas de cette partie du travail. Elle vous incombe exclusivement et est primordiale.

Les hypothèses :

Le détail des hypothèses est le suivant :

- Normalité des erreurs
- Homoscédasticité (égalité des variances des erreurs)
- Indépendance des erreurs

L'indépendance des erreurs

Il n'existe pas de test permettant de déterminer si les erreurs sont indépendantes ou non.

Un test que vous pourrez rencontrer est le test de Durbin-Watson qui détermine s'il y a une corrélation temporelle entre les résidus. Une telle corrélation peut découler, par exemple, de l'utilisation d'un appareil de mesure qui se déréglerait progressivement.

Généralement une représentation graphique des résidus permet de « voir »si l'hypothèse est réaliste. Attention aux données appariées !

L'égalité des variances

Cette hypothèse est **primordiale.** En effet à la fois le test de Kruskall-Wallis (équivalent non-paramétrique de l'ANOVA) et l'ANOVA requièrent qu'elle soit vérifiée.

Il convient ainsi de la tester avant l'hypothèse de normalité de l'erreur. Puisque l'on ne connaît alors pas encore la loi des erreurs il faut utiliser un test **non-paramétrique.**

Il s'agit du test de Levenne qui est utilisable dès que le nombre de répétitions pour chaque niveau du facteur est supérieur ou égal à trois.

Si l'hypothèse d'homoscédasticité est vérifiée on peut passer à la vérification de l'hypothèse de normalité des erreurs.

Dans le cas contraire, Minitab ne propose pas de test prenant en compte ce défaut. Si par contre vous avez accès à d'autres logiciels comme R, qui est gratuit et disponible sur internet mais assez difficile d'accès, SPSS, payant mais intuitif, ou SAS, pour les utilisateurs expérimentés uniquement, vous aurez à votre disposition des tests pouvant prendre en compte l'inégalité des variances.

La normalité des erreurs

Afin de tester la nomalité des erreurs on doit calculer ce que l'on appelle les résidus du modèle. En termes statistiques, il s'agit des réalisations des variables d'erreur $\epsilon_{i,j}$.

On doit donc commencer par calculer une estimation des coefficients $\mu, \alpha_1, \ldots, \alpha_I$ du modèle.

On notera toujours une estimation d'un coefficient c du modèle statistique par $\widehat{c}$.

Estimation

Il s'agit d'une valeur que l'on calcule à partir des observations de telle sorte que l'on juge qu'elle est une représentation de la valeur du paramètre c.
Par exemple vous connaissez une estimation de la moyenne μ d'une population par un échantillon d'effectif K :

$$\widehat{\mu} = \frac{1}{|K|} \sum_{k} x_k$$

En vous réferrant au cours de L3, vous êtes capables de caculer les estimations $\widehat{\mu}, \widehat{\alpha_1}, \ldots, \widehat{\alpha}_I$.

Les résidus

$$e_{i,j} = y_{i,j} - \mu - \alpha_i$$

Un résidu $e_{i,j}$ n'est rien d'autre que le défaut d'ajustement du modèle statistiqiue pour la $j$ème répétition du $i$ème niveau du facteur.

On doit alors tester la normalités des variables d'erreur. Or les effectifs ne permettent généralement pas de séparer les différents niveaux du facteur explicatif et de tester la normalité des résidus pour chacun des niveaux.

Pour obtenir une puissance convenable cela exigerait plus d'une cinquantaine de répétitions par niveaux !

On décide alors de regrouper tous les résidus ensemble et de procéder à **un** test de normalité sur tous les résidus.
Le nombre de tests de normalité existants est très important, il y a même des livres entiers sur le sujet... Lequel doit-on utiliser ?

Les logiciels de calcul statistique mettent à la disposition de l'utilisateur plusieurs tests de normalité. Bien entendu ils ont tous leurs qualités mais dans le contexte qui est le notre, c'est-à-dire celui de petits échantillons, effectif entre 10 et 100, c'est le test de Shapiro-Wilk qui est recommandé.

Tous les logiciels mentionnés plus haut permettent de réaliser ce test et en particulier Minitab, attention il a été renommé en test de Ryan-Joiner.

Si le test de normalité des résidus est significatif, vous n'avez pas le droit d'utiliser les résultats de l'ANOVA.

Il vous faut alors opter pour une alternative non paramétrique, le test de Kruskal-Wallis.

Ce test est disponible dans le logiciel Minitab.

Si la normalité n'est pas rejetée, on reteste l'hypothèse d'égalité des variances en utilisant cette fois le test de Bartlett.

Ce test est plus puissant que le test de Levenne car il repose sur une hypothèse de normalité des variables. C'est donc un test paramétrique.

Si l'hypothèse d'homoscédasticité n'est toujours pas rejetée, le modèle statistique est alors vérifé et l'on peut utiliser les résulats de l'ANOVA.