# Scénario 8

TABLEAUX A DEUX ENTRÉES

## 1 Présentation générale

### 1.1 Objectifs :

- Construire, interpréter, représenter graphiquement un tableau de mesures classifié par deux caractéristiques.
- Réaliser l'ajustement des données au modèle additif.

### 1.2 Les étapes

1. Construire et interpréter un tableau à deux entrées sur deux exemples (rendement de variété d'orge, production laitière).
2. Deux types de représentations graphiques mises en œuvre sur les mêmes exemples.
3. Permutations des lignes ou des colonnes du tableau de départ : leur intérêt.
4. Ajustement au modèle additif : comment trouver une formule simple pour calculer des valeurs approchées des mesures du tableau, faciles à interpréter.
5. Représentation graphique des données ajustées.
6. Comparaison entre un tableau de contingence et les tables étudiées dans ce scénario.
7. Evaluation (individuelle, anonyme).

### 1.3 Le matériel

- pour les stagiaires : crayons de couleur, compas, calculette, 2 feuilles de papier quadrillé pour l'évaluation.
- pour le formateur : un tableau (assez grand), projecteur de transparents, feutres à transparents, de couleur, permanents.

### 1.4 La forme

- Travail des stagiaires sur un premier fascicule (Feuilles 1, 1bis, 2, 3 et 4) pour les étapes 1 et 2 avec correction collective (transparents).
- Transparents pour l'étape 3 (dont une copie est distribuée aux stagiaires) (Transparents 5, 6, 7, 8, 5bis)
- Travail des stagiaires sur le 2ème fascicule pour l'étape 4 (Feuilles 9, 10, 11, 12). Bilan collectif.
- Discussion collective pour l'étape 5, support tableau
- Evaluation individuelle (feuilles 13, 14, 15).

### 1.5 Evaluation

Une évaluation individuelle sur place (sur feuille no 2 de l'étape 1).

Evaluation individuelle en dernière étape ou après la séance.

au choix : 3 tableaux à trous à remplir et un graphique à terminer

ou un tableau 2 × 2 à étudier selon la démarche du scénario (sous une forme simplifiée, guidée par 4 questions).

### 1.6 Prérequis

Calcul de moyennes, d'écarts. Graphiques de points repérés par deux axes : Facteur, variable (scénario 2).

### 1.7 Durée

Une séance. Evaluation finale éventuellement quelques jours après ou à la séance suivante.

## 2 Les étapes

### 2.1 Etape 1 : Construire et interpréter un tableau de mesures à deux entrées

- Faire construire le tableau des rendements de 5 variétés d'orge sur 6 stations du Minnesota (Chambers et al., page 384)

  Feuille stagiaire no 1. Transparent no 1 pour contrôle.

- Distribuer le tableau 2 (Feuille stagiaire no 1bis). Il s'agit du rendement laitier de vaches de 3 races dont le régime est complété en protéïnes avec 3 doses différentes. Il peut s'agir d'un plan d'expérience avec 9 vaches, ou avec $9r$ vaches chaque case représentant la moyenne de la production de $r$ vaches. (Il est important que ce nombre $r$ soit le même, sinon la précision de chaque mesure n'est pas la même, ce qui rend impossible les interprétations ultérieures).

  Chaque stagiaire répond aux questions sur sa feuille, réponse collective ensuite.

- Evaluation : à partir des données Ficus (données du scénario no 2) de la feuille stagiaire no 2, construire le tableau des accroissements des plantes, classifié par la température du sol et la position Est/Ouest (Transparent no 2 pour contrôle).

### 2.2 Etape 2 : Représentations graphiques

- Retour aux productions laitières (feuille no 2).

  Comment faire une représentation graphique?

  Deux solutions : la plus proche du tableau, consiste à dessiner une tâche (ronde ou carrée par exemple) à la place du chiffre, de rayon (ou de surface) proportionnelle à la mesure. Feuille stagiaire no 3, Transparent no 3.

  La deuxième à faire un axe avec la production et un axe avec soit la dose de soja, soit la race.

  Faire remarquer que la façon de positionner la race est *arbitraire*. Feuille stagiaire 4 et transparent 4.

  Interpréter : la meilleure race, celle qui profite le mieux d'une augmentation de la dose ...

- Montrer les représentations équivalentes pour les données *orge* de la feuille 1. Transparents 5,6,7 (les colorier avant usage, une couleur par variété, une par station).

  Question à poser : Quelle est la meilleure variété? (Trebi sauf pour la station 5 . . . ), quelle est la meilleure station? la meilleure combinaison? quelle est la bonne question à se poser pour un agriculteur précis?

## 2.3 Etape 3 : Permutations de lignes ou de colonnes.

Pour les transparents 6 et 7, remarquer encore que le choix des abscisses, de l'ordre des stations ou des variétés est arbitraire. Découper le transparent 7bis en 6 colonnes après l'avoir colorié et échanger devant eux les colonnes. Montrer qu'on peut choisir un ordre qui rend croissante une ligne brisée donnée (la variété Trebi par exemple, de façon générale on choisit celle qui joue le rôle de témoin)

Montrer éventuellement qu'en écartant plus ou moins les bandes, on peut faire une droite de la ligne *Trebi*. (Transparent 8). Cela permet de comparer facilement une variété quelconque à la variété Trebi.

**Remarque pour le formateur :** on a choisi comme abscisse de chaque station, le rendement de Trebi à cette station. La ligne Trebi devient donc la première bissectrice.

Dire que, en l'absence de témoin, on peut ordonner les stations sur l'axe des $x$ par rapport à leur rendement moyen pour les 5 variétés.
Quel effet sur le graphique en cercles du transparent 5 si on change l'ordre des stations? (on peut recopier les colonnes du transparent 5 sur des bouts séparés et les échanger devant eux). Si on change l'ordre des variétés? Faire les deux opérations implique d'avoir 30 petits bouts de transparents! Sur le transparent 5bis on a mis les variétés dans l'ordre des sommes par station croissantes et les stations dans l'ordre des sommes par variété croissantes.
Remarquer que cela change la lisibilité du schéma : s'il y a une tendance générale du genre "une station meilleure que les autres pour toutes les variétés", ça se voit ainsi que d'éventuelles exceptions à de telles tendances.

*Exercice facultatif :* recopier le tableau 1 (feuille 1) selon cet ordre. Faire peut être seulement remarquer que sur les chiffres eux-mêmes on peut faire les mêmes permutations et qu'on y voit déjà plus clair.

Pour mémoire (voir annexe)

sommes par ligne : M=1133 S=1093 V=1190 T=1420 P=1231

sommes par colonne : S1=928 S2=1362 S3=1053 S4=1090 S5=805 S6=829

ordre croissant : (S,M,V,P,T) et (5,6,1,3,4,2)

## 2.4 Etape 4 : Ajustement au modèle additif.

Retour aux données *production laitière* (feuille 4)

Remarquez que :

1. la production augmente avec la dose de soja
2. la production augmente si on passe de Normandie à Frisonne, et de Frisonne à Holstein.

On pourrait essayer de résumer la production d'une vache quelconque par une formule du type :

Production d'une vache = Production moyenne de la race + effet différentiel du régime.

Distribuer les feuilles 9,10 et 11.

Noter au tableau quelques résultats si besoin.

**Résultats :**

Moyenne générale = 7178
Moyennes N,F,H = 5567, 7167, 8800
Effet (1.8, 1.9, 2.0) = -311, 122, 189

Si tout va bien, conclure en précisant :

- que la somme des effets est nulle
- qu'on a écrit un modèle, appelé *modèle additif*
- que les valeurs observées ne s'ajustent pas exactement d'où le calcul des écarts ou résidus. Qu'ils ont un signe. L'interpréter.
- que si ces résidus sont petits (leur somme de carrés par exemple) , le modèle est bon, c'est-à-dire cohérent avec les données; le modèle s'ajuste bien aux données. On dit aussi que les données s'ajustent au modèle (mais c'est moins correct car c'est le modèle qui est éventuellement à remettre en cause, pas les données). Il est alors facile de résumer un grand tableau par peu de nombres. Lesquels? Combien?

  **Réponse :** nombre de lignes + nombre de colonne - 1, car la somme des effets est nulle.
- qu'on dit aussi *valeur estimée* par le modèle au lieu de *valeur ajustée*
- qu'un tel modèle, s'il est bon peut servir à faire de la prédiction : si une mesure ou l'autre est manquante, on peut l'estimer par ce que donne la formule : c'est peut être un peu faux, mais mieux que rien.

Réflexions pour le formateur

1. Il semble difficile de faire ici la distinction entre les paramètres du modèle, les $\mu_i$ et les $\beta_j$

   $$Y_{ij} = \mu_i + \beta_j + \varepsilon_{ij}$$

   et leurs estimations, qu'on vient de manipuler.

   De même, une comparaison détaillée avec la régression est difficile. On peut dire que choisir d'ajuster une droite pour chaque race de vaches (quand on met les doses en abscisses) et imposer que ces droites soient parallèles correspond à un modèle plus simple que le modèle additif, mais qui s'ajustera moins bien. La notion de nombre de paramètres du modèle peut être intuitivement présentée : pour la régression, avec deux nombres, la pente et l'ordonnée à l'origine on peut, pour un $x$ quelconque, calculer une valeur estimée; pour le modèle additif avec $l+c-1$ nombres on peut, pour tout niveau de chacun des facteurs, calculer une valeur estimée.

   Tout ceci est tout à fait facultatif et ne doit être envisagé que si les stagiaires sont prêts à le comprendre.

2. On peut aussi, surtout si quelqu'un y pense, écrire la formule sous l'autre forme :

   Production = Production moyenne pour la dose + effet différentiel de la race

   et , pour plus de symétrie,

   Production = Production moyenne + effet dose + effet race

   Mais ce n'est pas indispensable.

## 2.5 Etape 5 : Représentation graphique du modèle additif.

Sur la feuille 12, les stagiaires font les graphiques de valeurs ajustées calculées sur la feuille 11.
On obtient des lignes *parallèles* : cela caractérise le modèle additif. Interpréter : passer d'une race à l'autre donne la même différence, quelle que soit la dose de soja. De même augmenter la dose de 0,1 kg donne une même augmentation de la production, quelle que soit la race, ceci *si le modèle est exact* ou si l'approximation donnée par le modèle nous suffit.

Faire faire sur papier quadrillé, deux graphiques des résidus. (Ils remarquent qu'il faut placer la ligne zéro au milieu du dessin). Comparer ceux-ci aux précédents (feuilles 4 et 12).

## 2.6 Etape 6 : Tables de contingence, table à deux entrées.

On travaille collectivement (au tableau) en parallèle sur deux tables : celle des vaches (production laitière) et la suivante. Sur une population de 1469 hommes nés en 1920 on note au 31.XII.80 , pour chacun s'il est décédé ou non et s'il est (a été) fumeur ou non.

| | Fumeurs | Non Fumeurs |
|---|---|---|
| survie $>$ 60 ans | 348 | 950 |
| survie $\leq$ 60 ans | 54 | 117 |

Pour chaque tableau, répondre en parallèle aux questions suivantes :

1. on veut étudier l'influence de ...sur ...ou la relation entre ...et ...
2. nombre d'unités expérimentales (9 ou 1469)
3. nombre de facteurs contrôlés (2 ou 0)
4. nombre de variables (1 ou 2)
5. nature (continue, discrètes à 2 valeurs chacune)
6. construire (le début de) 2 tableaux avec une ligne par individu et une colonne par facteur ou par variable
7. le tableau de contingence résume le 2ème (sans perdre d'information autre que le no des individus), la table des vaches le recopie intégralement en le présentant autrement.

## 2.7 Etape 7 : Evaluation.

Deux possibilités, la 2ème demandant d'avoir mieux compris la cohérence du scénario.

- estimation des rendements de variétés d'orge sur 6 stations avec le modèle additif. Distribuer les feuilles 13 et 14, à compléter.

  *Remarque :* une critique des stagiaires parce que le tableau n'est pas ordonné sera appréciée

- faire un graphique des données, le calcul des valeurs ajustées au modèle additif, le graphique de ces valeurs ajustées sur le même papier pour un des deux exemples du scénario 2

  1. Agneaux : Mélange de Graminé x Pâture
  2. Moutarde : Variété x Lumière

  (Feuille no 15 , réponse sur papier libre, dont une feuille quadrillée pour les graphiques).

## 2.8 Mots-clé

Deux facteurs de classification des mesures, un par entrée
Permutation des lignes, permutation des colonnes, ligne croissante
Valeur ajustée = valeur estimée ou prédite par le modèle
Modèle = description simplifiée des données
Résidu négatif ou positif
Lignes parallèles (différence avec droites parallèles)
Approximation (des valeurs mesurées par les valeurs ajustées)

## 2.9 Liste du vocabulaire

Résidu (residual)
Tableau à deux entrées (two way table)
Ajustement ou valeur ajustée (adjusted value)
Modèle additif (additive model)
Modèle (mathématique)

## 2.10 Références bibliographiques

CHAMBERS J.M., CLEVELAND W.S., KLEINER B., EY P.A. (1983) *Graphical methods for data analysis.* Wadsworth.

# 3 Annexes

## 3.1 Documents pour les stagiaires

Description du scénario 8
Vocabulaire du scénario 8
Feuilles stagiaires 1 à 4
Copie réduite des transparents 5 à 8 (feuilles stagiaires 5 à 8)
Feuilles stagiaires 9 à 12
Feuilles stagiaires pour évaluation 13 à 15

**N.B.** Les numéros des feuilles et des transparents se correspondent. D'où des trous dans la numérotation des transparents.

## 3.2 Documents pour le formateur

Transparents 1 à 4, 5 à 8
Résultats de la feuille 13
Transparent 16 des résidus orge (évaluation)

## 3.3 Les transparents à projeter