Chapitre 3

Comparaison de deux moyennes

3.1 Données, notations, exemples, question

3.1.1 Présentation des données et notations

On dispose de deux séries numériques comportant respectivement n_1 et n_2 valeurs (souvent $n_1 = n_2 = n$). Chaque série est extraite d'un ensemble (= population) d'où la dénomination *échantillon*. Sur un tableur, on présente souvent les deux séries séparément mais les logiciels statistiques nécessitent de regrouper toutes les valeurs en faisant apparaître une variable n^o *de l'échantillon* (cf. Tableau 3.1)

Echantillon 1	
n^o	x_1
1	x_{11}
i	x_{1i}
n_1	x_{1n1}
moy.	$\bar{x}_1$
ec.-ty.	s_1

Echantillon 2	
n^o	x_2
1	x_{21}
l	x_{2l}
n_2	x_{2n2}
moy.	$\bar{x}_2$
ec.-ty.	s_2

	n^o éch.	x
1	1	x_{11}
i	1	x_{1i}
n_1	1	x_{1n1}
1	2	x_{21}
l	2	x_{2i}
n_2	2	x_{2n2}

TAB. 3.1 – Notations et présentation des données.

Exemple numérique Deux variétés de blé ont été cultivées chacune sur 8 parcelles ($n_1 = n_2$ = 8). Les rendements observés (en quintaux/hectare) sont regroupés dans le tableau 3.2 (transposé par rapport au tableau 3.1 pour des raisons de mise en page).

	Valeurs observées (en q/ha)								moy.	ec.-ty	s'
Echantillon 1	78,6	78,9	79,3	79,8	80,5	80,6	81	81,3	80,0	0,935	1,000
Echantillon 2	80	80,3	80,9	81,5	82	82,2	82,5	82,6	81,5	0,935	1,000

TAB. 3.2 – Exemple numérique : deux séries de 8 rendements de blé. A l'intérieur de chaque échantillon, les données ont été triées. Définition de s' : cf. § 2.3.3 page 26

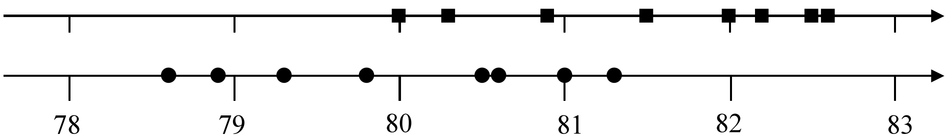

FIG. 3.1 – Représentations axiales des rendements en q/ha du tableau 3.2. Les données de l'échantillon 1 (resp. 2) sont représentées par des disques (resp. carrés)

Question posée Si l'on ne considère que les 16 parcelles, la variété 2 présente en moyenne un rendement supérieur (de 1,5 q/ha) à celui de la variété 1. Peut-on généraliser ce résultat ? Autrement dit, la différence observée (1,5 q/ha) doit-elle être considérée comme une conséquence d'un rendement moyen différent selon la variété ou, au contraire, est-il fortuit ? La question est légitime : même si les variétés ont exactement le même rendement moyen, il est très peu probable que l'on observe exactement le même rendement moyen dans une telle expérience. Selon un autre point de vue, la question peut être posée ainsi : la différence de moyennes observée doit-elle être imputée au hasard (c'est-à-dire à la variabilité « naturelle », dite aussi « résiduelle » pour exprimer que l'on ne sait l'expliquer par la statistique) ? Pour cela, il est nécessaire de disposer au moins d'un ordre de grandeur de cette variabilité, soit par une connaissance a priori (situation rencontrée lorsque l'on réalise souvent un même type d'expérience) soit à partir des données elles-mêmes (la variance des valeurs d'un même échantillon en donne une estimation).

La procédure classique de test statistique montre comment confronter une différence de moyennes observée à une variabilité « naturelle » en prenant en compte les risques d'erreur. Ses différentes variantes dépendent des informations disponibles sur la variabilité.

Autres exemples Une société de vente par correspondance compare deux types d'offres commerciales. Elle adresse chacune des offres à un échantillon de 1000 clients. A chaque client est associé le chiffre d'affaire engendré par l'offre qu'il a reçue. Les deux offres sont–elles équivalentes ?

Les hommes et les femmes sont-ils également sensibles à la saveur sucrée ? Dans la perspective de cette question, on a mesuré le seuil de perception de la saveur sucrée de 20