

Statistiques descriptives bivariées

Objectifs

- Observer simultanément des individus d'une population sur deux caractères
- Mesurer un lien éventuel entre deux caractères en utilisant un résumé chiffré qui traduit l'importance de ce lien
- Qualifier ce lien :
 - en cherchant une relation numérique approchée entre deux caractères quantitatifs
 - en cherchant des correspondances entre les modalités de deux caractères qualitatifs

2 types de variables $\Rightarrow$ 3 types de croisements :

- qualitatif $\times$ qualitatif
- qualitatif $\times$ quantitatif
- quantitatif $\times$ quantitatif

Croisement qualitatif $\times$ quantitatif

Croisement Qualitatif × Quantitatif

Pour étudier le lien entre une variable qualitative à p modalités et un caractère quantitatif, on partitionne la population P en sous-populations : une sous-population pour chaque modalité de la variable qualitative

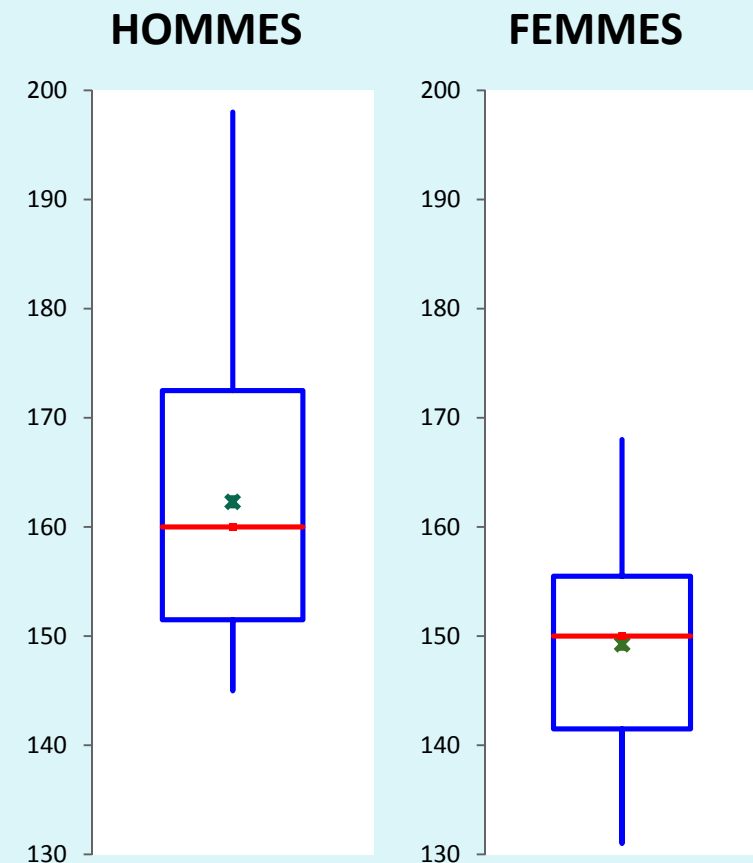
Prenons l'exemple de l'étude comparative de la taille entre les hommes et les femmes.

La variable qualitative est le sexe. On divise donc l'échantillon en deux sous-échantillon, celui des hommes et celui des femmes.

Sur chaque sous-échantillon, il est possible de calculer les résumés numériques usuels.

Une comparaison des boîtes de Tukey permet d'avoir une première idée du lien entre les variables

SEXE	Effectifs	Moyenne	Écart-type
Hommes	23	162,30	14,21
Femmes	35	149,29	10,52
Total	58	154,45	13,61

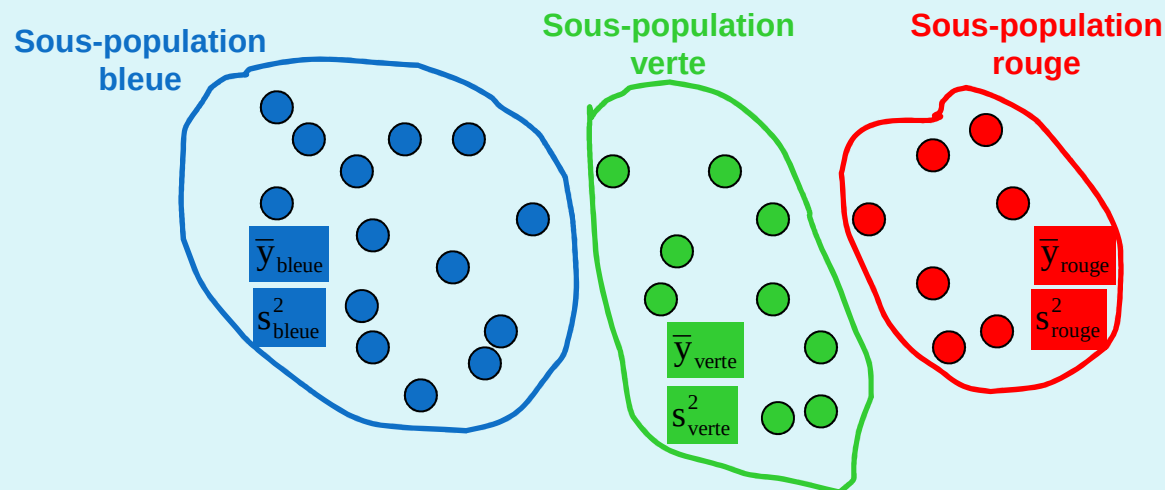


Croisement Qualitatif × Quantitatif

Variance intra

On étudie la variable quantitative Y sur chaque sous-population en calculant la moyenne et la variance de Y . On parle de *variation intra*,

$$\text{var}^{\text{intra}}(Y) = \frac{1}{n} \sum_{l=1}^p \overbrace{n_l}^{\text{Effectif de la sous-population } \ell} \underbrace{s_l^2}_{\text{Variance de la sous-population } \ell} = \frac{1}{n} \sum_{l=1}^p n_l \times \frac{1}{n_l} \sum_{i=1}^{n_l} (y_i - \overbrace{\bar{y}_l}^{\text{Moyenne de la sous-population } \ell})^2$$



Décomposition de la moyenne

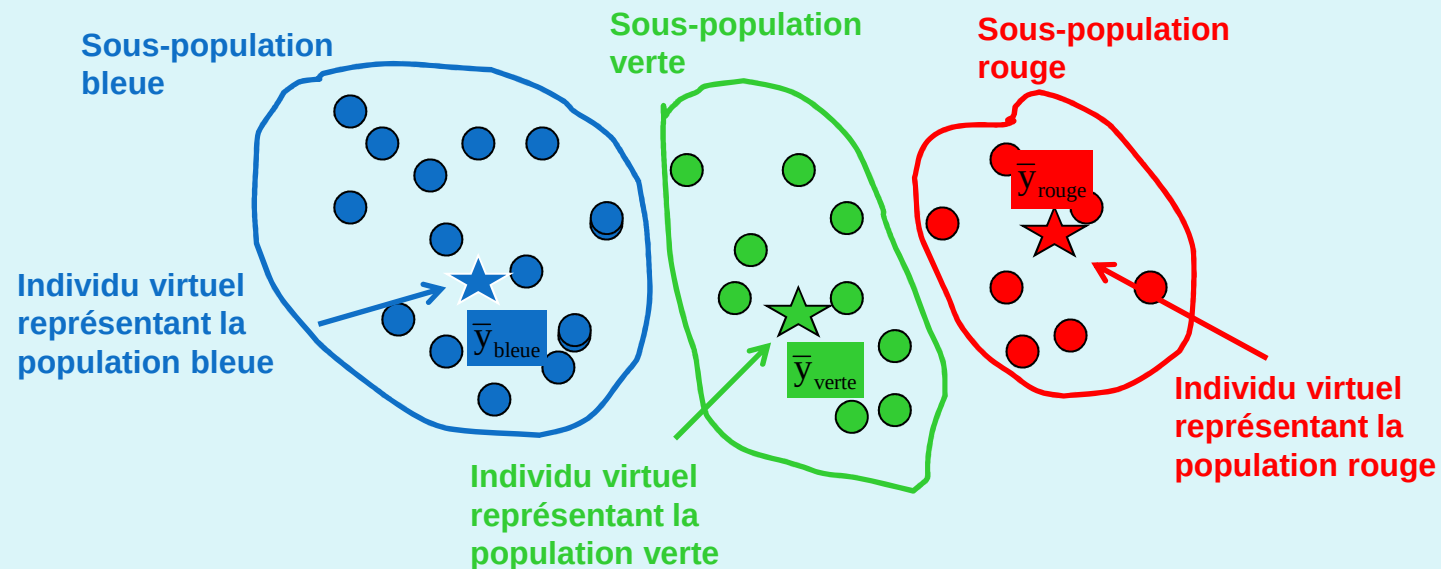
$$\frac{1}{n} \sum_{\ell=1}^p n_{\ell} \bar{y}_{\ell} = \bar{y}$$

Croisement Qualitatif × Quantitatif

Variance inter

- Pour chaque sous-population, on crée un individu virtuel dont la valeur sur Y est égale à la moyenne des valeurs de Y des individus de la sous-population.
- On crée donc une nouvelle population formée de ces individus virtuels. Chaque individu aura un poids de n_i , l'effectif de chaque sous-population. On parle de *variation inter*,

$$\text{var}^{\text{inter}}(Y) = \frac{1}{n} \sum_{\ell=1}^p n_{\ell} (\bar{y}_{\ell} - \bar{y})^2$$



Croisement Qualitatif × Quantitatif

Rapport de corrélation

On peut donc définir trois variances sur variable Y.

- une première qui explique les variations de Y dans toute la population : totale
- une deuxième qui explique les variations de Y dans les sous- populations : intra
- une troisième qui explique les variations de Y entre les sous-populations : inter

Nous avons la décomposition de la variance suivante :

$$\text{var}^{\text{totale}}(Y) = \text{var}^{\text{inter}}(Y) + \text{var}^{\text{intra}}(Y)$$

Variance expliquée
Variance résiduelle

On en déduit une mesure du lien entre X et Y avec le *rapport de corrélation*

$$\frac{\text{var}^{\text{inter}}(Y)}{\text{var}^{\text{totale}}(Y)}$$

Le rapport de corrélation représente le pourcentage de variabilité de Y expliquée par X. Il varie entre 0 et 1. Si

- *S'il est nul alors la variance expliquée est nulle, il n'y a donc aucun lien entre Y et X*
- *S'il vaut 1 alors la variance expliquée est égale à la variance de Y donc Y est entièrement expliquée par X.*

Croisement Qualitatif × Quantitatif

Exemple

Etude comparative de la taille entre les hommes et les femmes.

SEXE	Effectifs	Moyenne	Variance	Var intra
Hommes	23	162,0	193,3	$23 \times 193,3 = 4444,9$
Femmes	35	149,3	107,6	$35 \times 107,6 = 3765,1$
Total	58	154,4	182,1	$(4444,9 + 3765,1) / 58 = 141,5$

SEXE	Effectifs	Moyenne	Variance	Var inter
Hommes	23	162,3	193,3	$23 \times (162,3 - 154,4)^2 = 1419,5$
Femmes	35	149,3	107,6	$35 \times (149,3 - 154,4)^2 = 932,8$
Total	58	154,4	182,1	$(1419,5 + 932,8) / 58 = 40,6$

On vérifie la formule de la décomposition de la variance : $141,5 + 40,6 = 182,1$

Rapport de corrélation :

$$\frac{40,6}{182,1} = 0,22$$

22% de la variabilité de la taille est expliquée par le sexe