Data exploration

Statistiques descriptives bivariées

- 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 ⇒ 3 types de croisements :

- qualitatif ×qualitatif
- **qualitatif ×quantitatif**
- quantitatif ×quantitatif

Croisement Quantitatif - Qualitatif

Sous-populations

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.

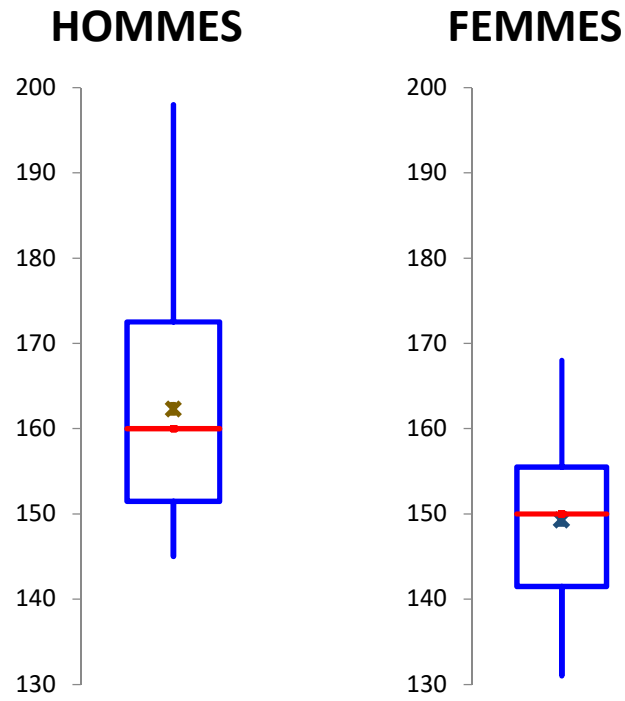
Prenons l'exemple de l'étude comparative de la taille entre les hommes et les femmes.

La variable qualitative est le genre. On divise donc la population en deux sous-populations, celui des hommes et celui des femmes.

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

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

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



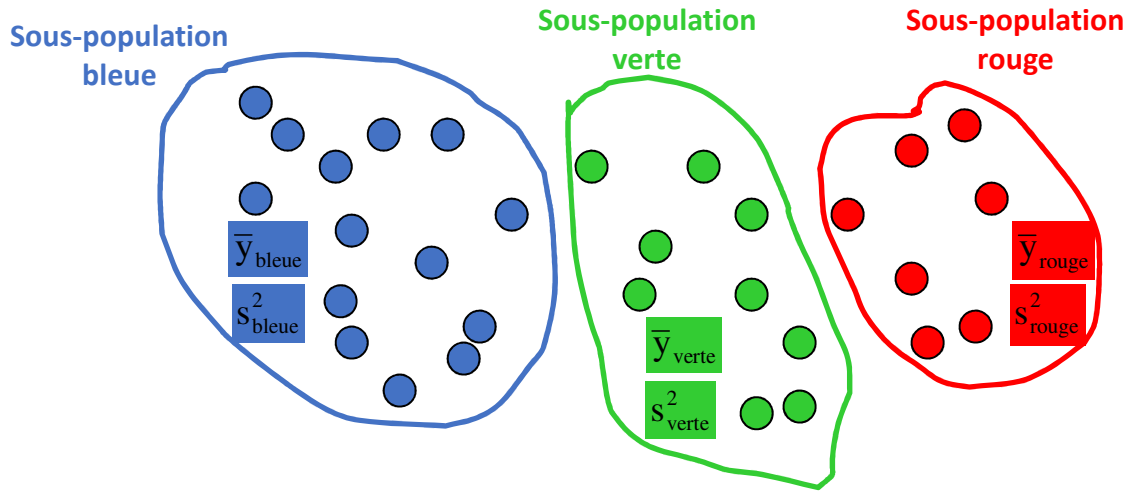
L'idée est de décomposer la moyenne et la variance sur les différents sous-populations.

Croisement Quantitatif - Qualitatif

Décomposition de la moyenne

Notons Y la variable quantitative et pour k=1,…,p

- n_k l'effectif de la kème sous-population
- $\bar{y}_k$ la moyenne de Y sur la kème sous-population
- s_k^2 la variance de Y sur la kème sous-population



La moyenne de la population est égale à la moyenne pondérée des moyennes des sous-populations,

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{p} n_k\,\bar{y}_k$$

Croisement Quantitatif - Qualitatif

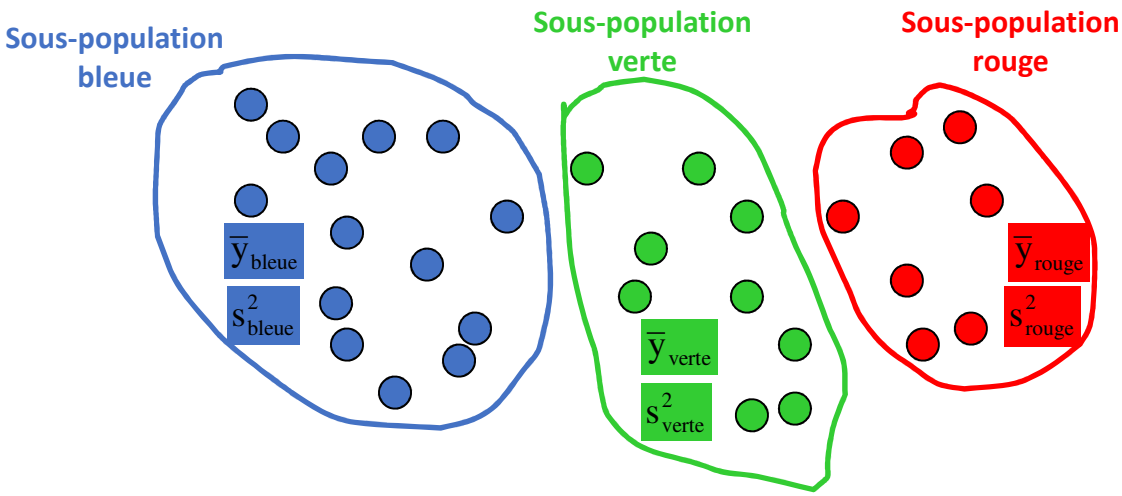
Décomposition de la variance

Contrairement à la moyenne, la variance de la population n'est pas égale à la moyenne pondérée des variances des sous-populations. Cette dernière s'appelle la **variance intra groupes**,

$$var^{intra}(\mathrm{Y}) = \frac{1}{n}\sum_{k=1}^{p} n_k s_k^2$$

Effectif de la sous-population k

Variance de la sous-population k



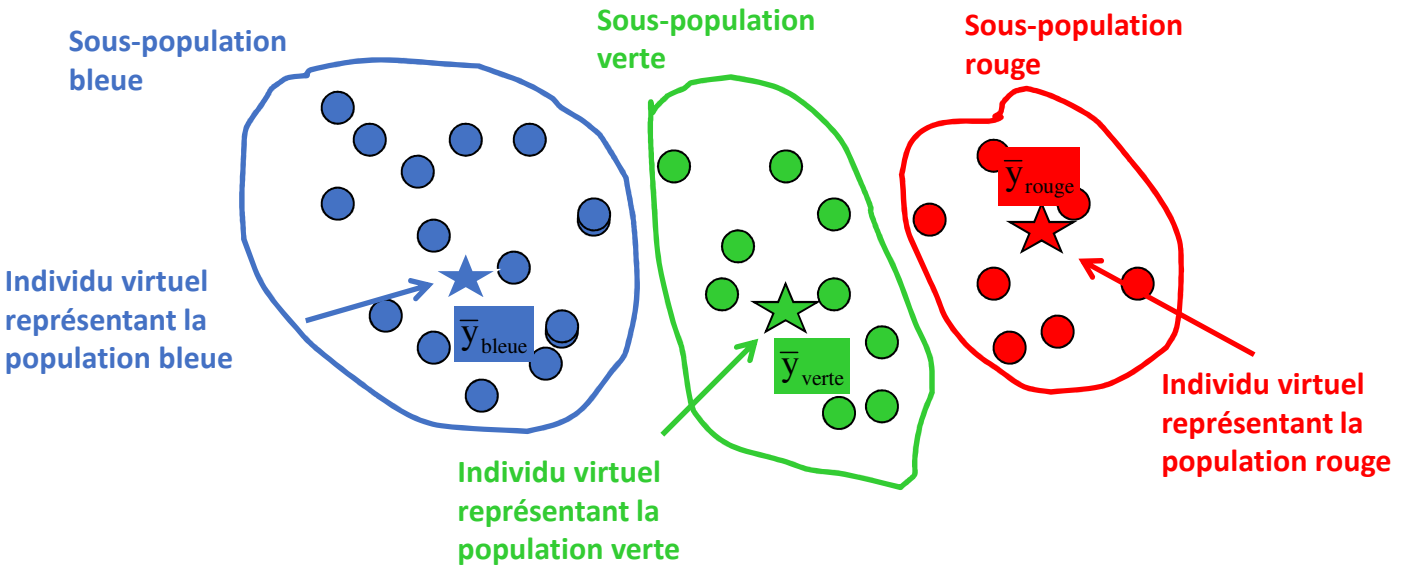
Croisement Quantitatif - Qualitatif

Décomposition de la variance

A la variance intra groupes, on ajoute la **variance inter groupes**,

$$var^{inter}(\mathrm{Y}) = \frac{1}{n}\sum_{k=1}^{p} n_k(\bar{y}_k - \bar{y})^2$$

- 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_k , l'effectif de chaque sous-population.



Croisement Quantitatif - Qualitatif

Décomposition de la variance

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 :

$$var^{tot}(Y) = var^{inter}(Y) + var^{intra}(Y)$$

Variance expliquée **Variance résiduelle**

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

$$\frac{var^{inter}(Y)}{var^{tot}(Y)}$$

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 Quantitatif - Qualitatif

Exemple

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

GENRE	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**

GENRE	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 genre