LIEN ENTRE UNE VARIABLE QUALITATIVE (e.g. REGION) ET UNE VARIABLE QUANTITATIVE (e.g. PRIX D'UN PRODUIT)

P : le prix (en €) est une variable C°.

R : la région a 3 modalités : IdF, SO et L-R

IdF : Ile-de-France ; SO : Sud-Ouest ; L-R : Languedoc-R

Tableau de données

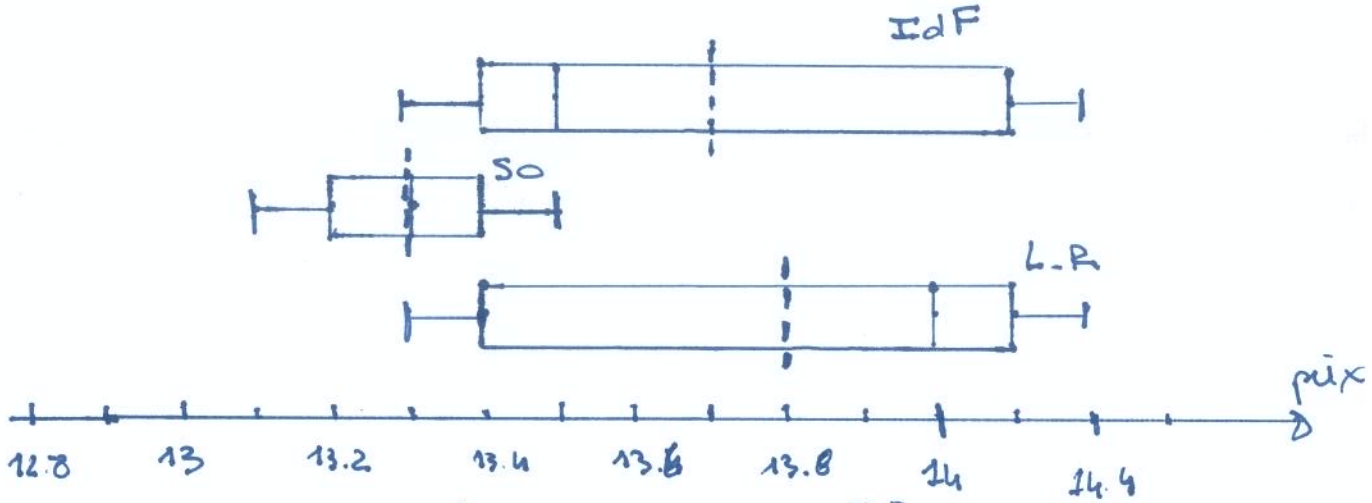
Observations	P	R
1	13.5	IdF
2	13.2	SO
3	13.4	L-R
4	14.2	IdF
5	13.3	SO
6	13.3	L-R
7	14.1	IdF
8	13.1	SO
9	14.0	L-R
10	13.4	IdF
11	13.5	SO
12	14.2	L-R
13	13.3	IdF
14	13.4	SO
15	14.1	L-R

Distributions conditionnelles à la région

Prix	Région		
	IdF	SO	L-R
Moyen	13.7	13.3	13.8
$Q_{25\%}$	13.4	13.2	13.4
Median	13.5	13.3	14
$Q_{75\%}$	14.1	13.4	14.1
Variance	0.14	0.02	0.14

(.../...)

On peut représenter ces distributions conditionnelles :



⇒ Il y a bien une disparité prix / région.

On va décomposer la variance totale

$$s_p^2 := \frac{1}{n}\sum (p_i - \bar{p})^2 \quad \text{où} \quad \bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

$$\left(= \underbrace{\frac{1}{n_{IdF}} \sum_{j \in IdF} p_j}_{= \bar{p}_{IdF}} + \underbrace{\frac{1}{n_{So}} \sum_{j \in So} p_j}_{= \bar{p}_{So}} + \dots \right.$$

en une variance intra-région et une variance inter-région

$$V_{intra} = \frac{1}{n} \sum_{"j" \in \{IdF, So, L-R\}} n_{"j"} s^2_{"j"} \quad \text{où} \quad s^2_{"j"} = \frac{1}{n_{"j"}} \sum_{i \in "j"} (p_i - \bar{p}_{"j"})^2$$

est la variance de la région "j"

et $n_{"j"}$ est l'effectif des observations de la région "j"

$$\text{et } V_{inter} = \frac{1}{n} \sum_{"j" \in \{IdF, So, L-R\}} n_{"j"} (\bar{p}_{"j"} - \bar{p})^2$$

Dans notre exemple numérique : $V_{inter} = 0.04667$

$V_{intra} = 0.1$

$s_p^2 = 0.14667$

Le rapport de corrélation est $\eta_{P|R} := \sqrt{\frac{V_{inter}}{s_p^2}} = \sqrt{1 - \frac{V_{intra}}{s_p^2}}$

Plus ce rapport est proche de 1, plus [...]

les groupes sont hétérogènes et donc il y a une liaison forte entre la variable qualitative (Région) et quantitative (Prix). Si $\eta_{P/R}$ est proche de 0, les moyennes conditionnelles fluctuent peu par rapport à la moyenne globale : peu de lien entre P et R.

→ Dans notre application numérique, $\eta_{P/R} = 0.564$ permet de conclure qu'il y a une disparité selon les régions ; l'examen des graphiques "boîtes à moustache" montre bien que c'est la région SO qui est franchement différente des 2 autres.
On notera au passage l'asymétrie (moyenne très différente de la médiane) des distributions conditionnelles des prix en IdF et L-R

---

(4) Démo de $S_P^2 = V_{intra} + V_{inter}$:

$$S_P^2 = \frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p})^2 = \frac{1}{n}\left[\sum_{i\in IdF}(p_i-\bar{p})^2 + \sum_{i\notin SO}(p_i-\bar{p})^2 + \sum_{i\notin L\text{-}R}(p_i-\bar{p})^2\right]$$

$(p_i-\bar{p}) = (p_i - \bar{p}_{IdF}) + (\bar{p}_{IdF} - \bar{p})$; $(p_i-\bar{p})^2 = \left[(p_i-\bar{p}_{IdF}) + (\bar{p}_{IdF}-\bar{p})\right]^2$ = idem

$\forall$ "j" $\in \{IdF, SO, L\text{-}R\}$

$$\sum_{i\in \text{"j"}}(p_i-\bar{p})^2 = \sum_{i\in \text{"j"}}\left[(p_i-\bar{p}_{\text{"j"}})^2 + (\bar{p}_{\text{"j"}}-\bar{p})^2 + 2(p_i-\bar{p}_{\text{"j"}})(\bar{p}_{\text{"j"}}-\bar{p})\right]$$

$$= n_{\text{"j"}}\cdot\underbrace{\frac{1}{n_{\text{"j"}}}\sum_{i\in \text{"j"}}(p_i-\bar{p}_{\text{"j"}})^2}_{=s^2_{\text{"j"}}} + n_{\text{"j"}}(\bar{p}_{\text{"j"}}-\bar{p})^2 + 0$$

, car $\sum_{i\in \text{"j"}}(p_i-\bar{p}_{\text{"j"}}) = 0$

En remplaçant les éléments : $= V_{intra}$ $\quad = V_{inter}$

$$S_P^2 = \frac{1}{n}\sum_{\text{"j"}} n_{\text{"j"}}\cdot s^2_{\text{"j"}} + \frac{1}{n}\sum n_{\text{"j"}}(\bar{p}_{\text{"j"}}-\bar{p})^2$$

□