Statistique pour l'ingénieur

Cours n° 3

Statistique descriptive bidimensionnelle

Statistique descriptive

- Etapes à suivre :
  - Décrire séparément chacune des variables observées *sur un même échantillon* ⇒ Description unidimensionnelle des données
  - Étudier simultanément les variables observées *sur un même échantillon* ⇒ Description bi(multi)dimensionnelle des données
  - Étudier les liaisons entre les variables observées ⇒ Etude des corrélations

- Les méthodes seront différentes selon la nature des variables observées.
  - 3 cas en 2D :
    (quantitatif, quantitatif), (quantitatif, qualitatif), (qualitatif, qualitatif)

quantitative vs. quantitative

- Ex: age vs. note au médian

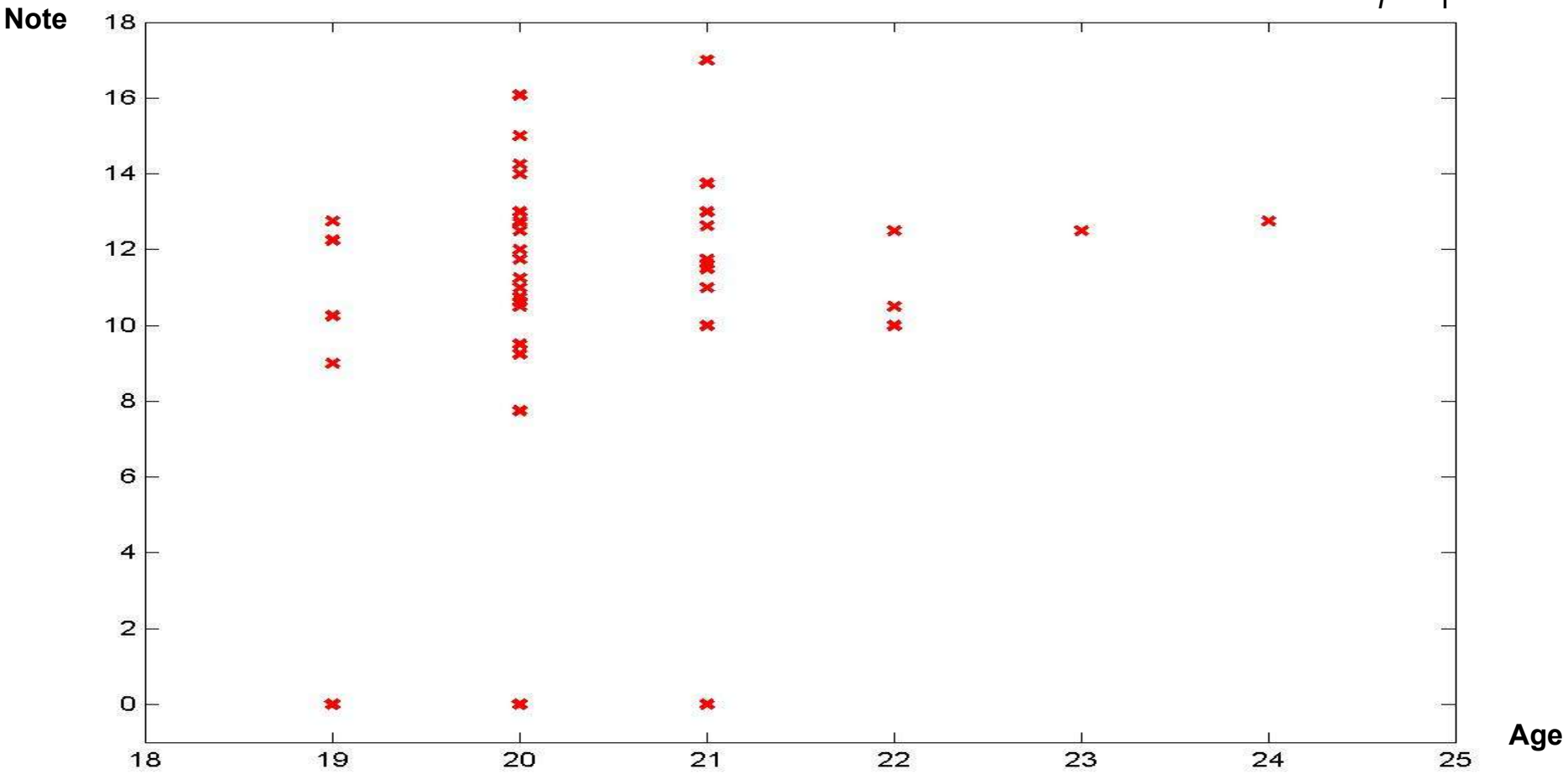
$$x^t = [x_1\ x_2 \ldots x_n]^t$$

$$y^t = [y_1\ y_2 \ldots y_n]^t$$

Age	note
19	12.25
19	9.0
20	12.5
...	
20	11.25
22	10.0
20	13.0
20	11.0
20	12.5
19	0
22	10.5
21	17.0
21	12.63

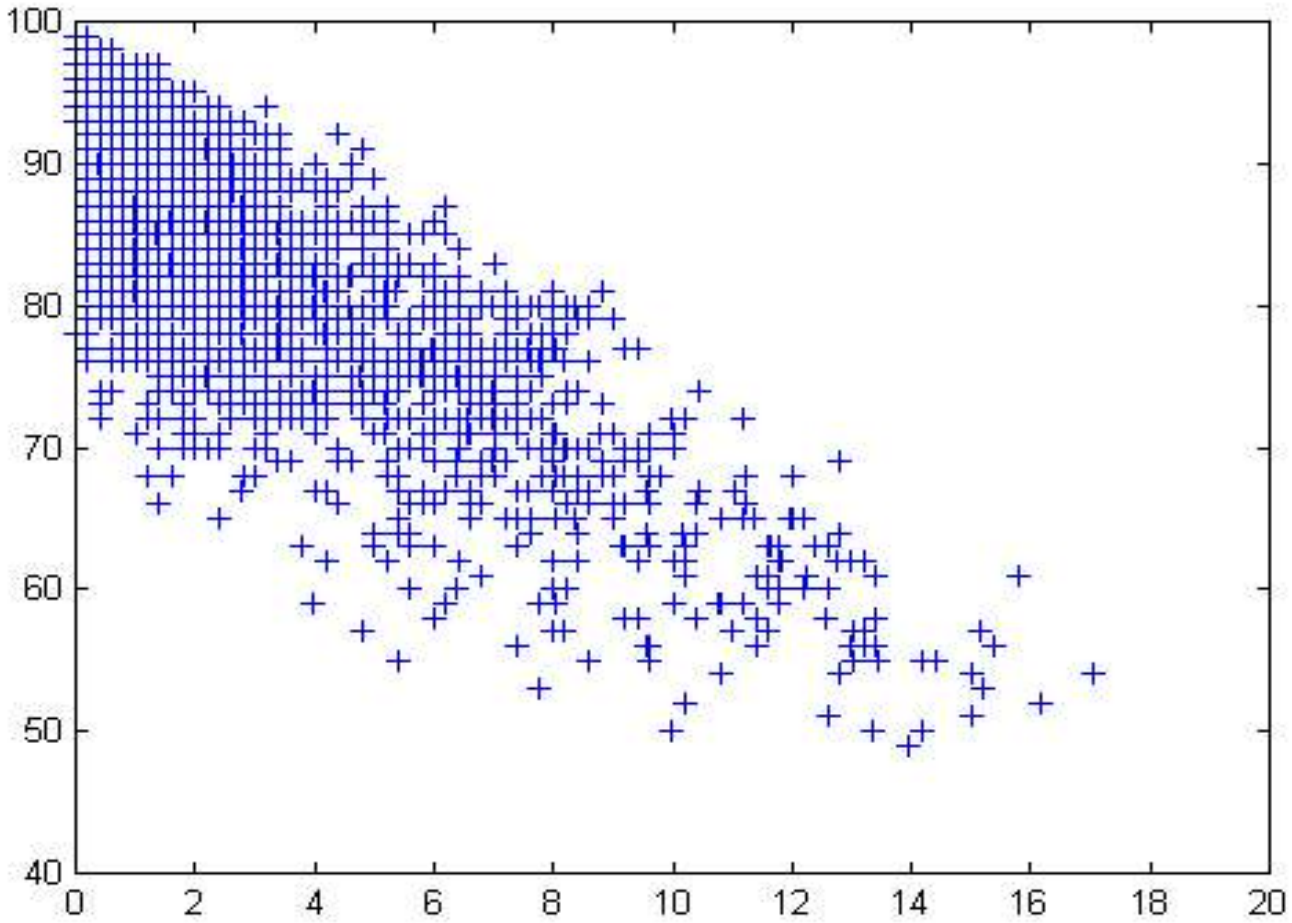
quantitative vs. quantitative

- Représentation graphique : le nuage de points $\{(x_i, y_i)\}$



quantitative vs. quantitative

- Autre exemple:
  - Y=usr - Portion of time (%) that cpus run in user mode
  - x_6=fork - Number of system fork calls per second



quantitative vs. quantitative

- Notion de corrélation

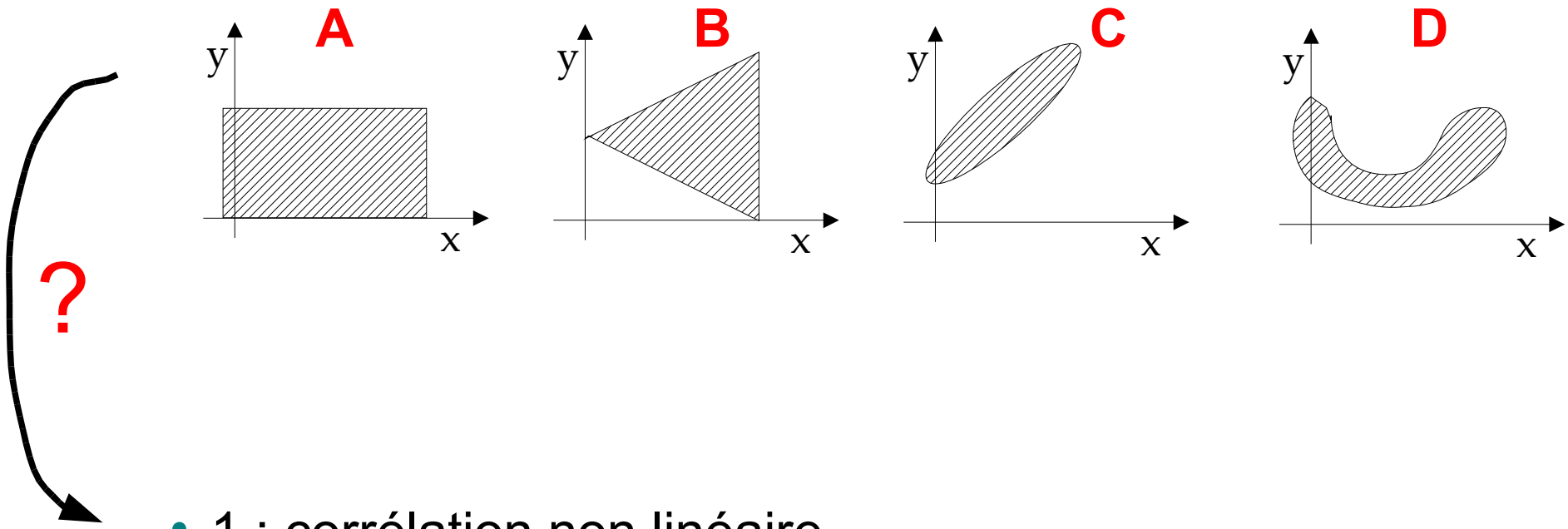
  - Corrélation si dépendance en moyenne :

    à $X=x$ fixé, la moyenne $\overline{Y}$ est une fonction de x

    - Remarque :
      non-corrélation ne signifie pas nécessairement indépendance

quantitative vs. quantitative

- Notion de corrélation



- 1 : corrélation non linéaire
- 2 : absence de liaison en moyenne mais pas en dispersion
- 3 : corrélation linéaire
- 4 : absence de liaison

quantitative vs. quantitative

- Coefficient de corrélation linéaire r_{XY}
  ⇒ mesure le caractère linéaire du nuage de points

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_X s_Y} = \frac{cov(X,Y)}{s_X s_Y} \quad avec \quad -1 \leqslant r \leqslant 1$$

où $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ et $s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$ sont des variances

- Propriété :

  si $|r| = 1$
  alors il existe une relation linéaire exacte $ax_i + by_i + c = 0 \ \forall i$

quantitative vs. quantitative

- Régression linéaire entre 2 v. a. quantitatives *X* et *Y*
  - Si *X* et *Y* sont « correctement » corrélées (r_{XY} voisin de 1)
    Si *X* est a priori la cause de *Y* ,

> Réaliser la régression de *Y* sur *X*
> = déterminer la fonction affine $f(Y)=aX+b$ qui approche « le mieux possible » *Y*

  - La détermination de *a* et *b* se fait selon le critère des moindres carrés, i.e. en minimisant :

$$S(a,b)=\sum_{i=1}^{n}\{y_i-[ax_i+b]\}^2$$

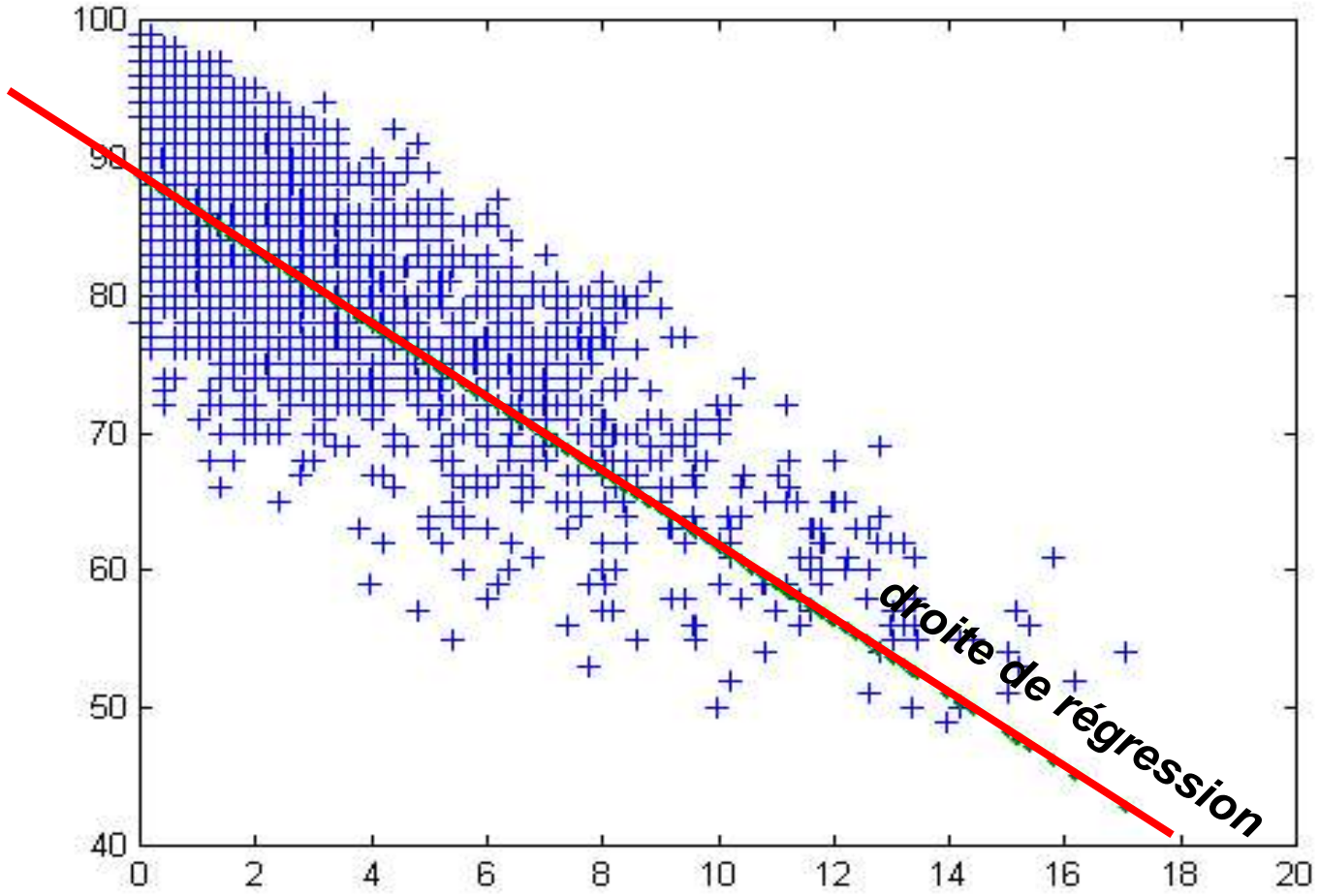
$$\Rightarrow \quad a^o=\frac{cov(X,Y)}{\sigma_X^2} \quad \text{et} \quad b^o=\overline{y}-a^o\overline{x}$$

quantitative vs. quantitative

- Régression linéaire entre 2 v. a. quantitatives *X* et *Y*
  - La droite de régression de *Y* sur *X* : $y=a^o x+b^o$ passe par le barycentre (de cordonnées $\overline{x}$ et $\overline{y}$) du nuage.
  - Les valeurs ajustées $y_i^o=a^o x_i+b^o$ ont la même moyenne que *Y*
  - Les résidus $e_i^o = y_i - y_i^o$ sont de moyenne nulle et de variance $S(a^o,b^o)/n$
  - La v. a. causale *X* et la v. a. résiduelle E^o sont non corrélées

quantitative vs. quantitative

- Exemple: la meilleure droite : $y = -2.6952\ x_6 + 88.7097$



quantitative vs. qualitative

- Soit *X* la variable qualitative à *r* modalités : $x^{(1)}$ ... $x^{(r)}$
- Soit Y la variable quantitative de moyenne $\bar{y}$ et de variance s_y^2
- X et Y sont observées sur un même échantillon Ω
  - Chaque modalité définit une sous-population Ω_λ de Ω avec $\lambda=1,...,r$, c.a.d. une partition de Ω en *r* classes

$\Rightarrow$ moyenne et variance partielles de *Y* sur chaque sous-population, notées $\bar{y}_\lambda$ et s_λ^2

quantitative vs. qualitative

- Formules de décomposition de la moyenne et de la variance de la variable quantitative Y sur la partition définie par la variable qualitative X :
  - $\overline{y} = \frac{1}{n} \sum_{\lambda=1}^{r} n_\lambda \overline{y_\lambda}$
  - $s_Y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n} \sum_{\lambda=1}^{r} n_\lambda (\overline{y_\lambda} - \overline{y})^2 + \frac{1}{n} \sum_{\lambda=1}^{r} n_\lambda s_\lambda^2 = s_E^2 + s_R^2$

  Où s_E^2 variance expliquée et s_R^2 variance résiduelle

quantitative vs. qualitative

- Indice : rapport de corrélation
  mesure la liaison entre 2 variables *X* et *Y*

  - $s_{\frac{Y}{X}} = \sqrt{\frac{s_E^2}{s_Y^2}}$ $avec\ 0 \leqslant s_{\frac{Y}{X}} \leqslant 1$

    - où s_E^2 variance de la variable quantitative *Y* expliquée par la variable qualitative *X*

    - et s_Y^2 variance de la variable quantitative *Y*

quantitative vs. qualitative

- Exemple
  - v.a. qualitative : sexe
    - $X^{(1)}$ = H $X^{(2)}$ = F
  - v.a. quantitative : note
    - $\overline{y}$ = 11.2
    - S_y^2 = 12.25 (S_y=3.5)
  - $\overline{y_1}$ = 11.1 $\overline{y_2}$ = 11.8

    S_y^2 = 14.4 S_y^2 = 2.4

Note	Sexe
12.25	H
9.0	F
12.5	H
0.0	H
9.5	H
...	
17.0	H
12.63	H

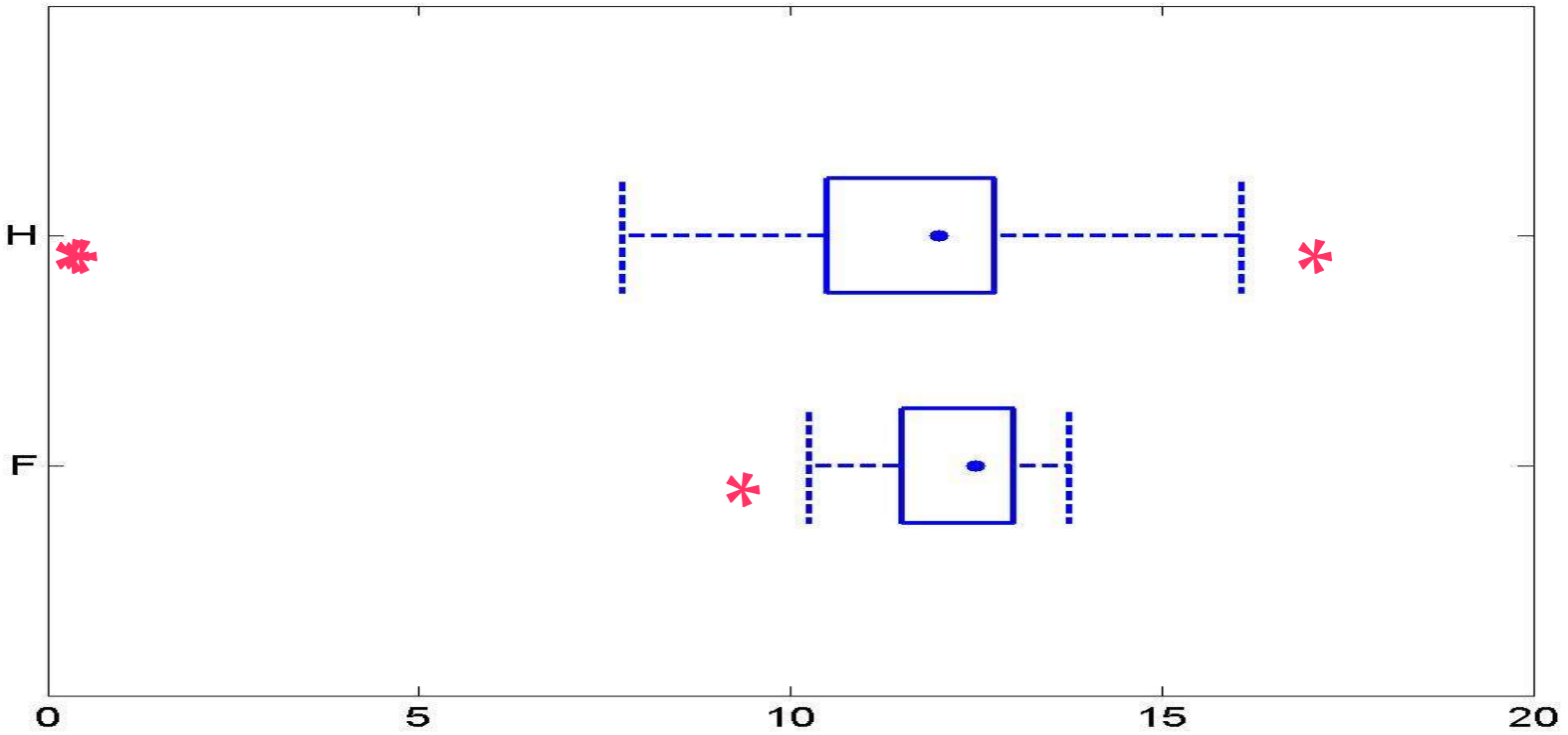
quantitative vs. qualitative

- $n^{(1)} = 37$ $n^{(2)} = 8$
- Décomposition de la moyenne
  - $45 * 11.2 = 37 * 11.1 + 8 * 11.8$
- Décomposition de la variance
  - $S_E^2 = [37 * (11.2-11.1)^2 + 8 * (11.2-11.8)^2] /45 = 0.07$
  - $S_R^2 = [37* 14.4 + 8 * 2.4]/45 = 12.2$
  - $S_y^2 = S_E^2 + S_R^2$
  - $S_{Y|X} = (0.07/12.25)^{1/2} = 0.08$ $\Rightarrow$ X et Y ne sont pas corrélés

quantitative vs. qualitative

Représentation graphique: boîtes parallèles

- sur un même graphique doté d'une échelle unique
- représenter pour *Y* une « boîte à moustaches » pour chaque sous-population définie par *X*



qualitative vs. qualitative

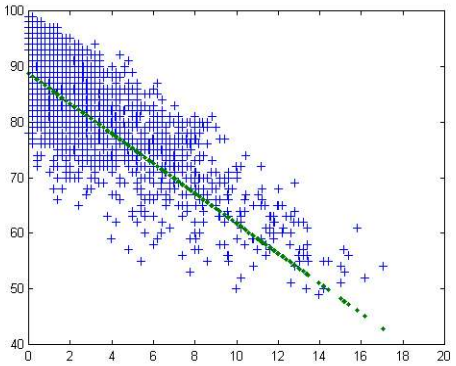
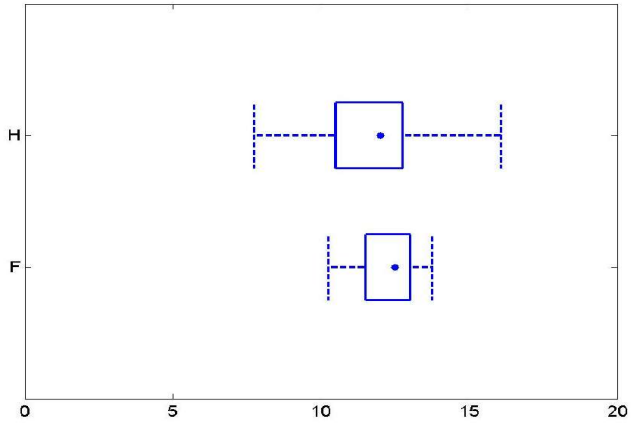
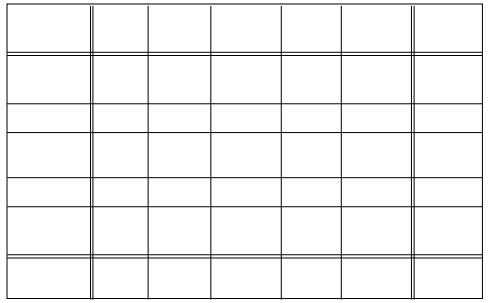
- Soit *X* une variable qualitative à *r* modalités : $x^{(1)} \ldots x^{(r)}$
- Soit Y une variable qualitative à *c* modalités : $y^{(1)} \ldots y^{(c)}$
- X et Y sont observées sur un même échantillon Ω
- Représentation : table de contingence

	$y^{(1)}$	…	$y^{(i)}$	…	$y^{(c)}$	Σ
$x^{(1)}$	n_{11}	…	n_{1i}	…	n_{1c}	n_{1*}
…						…
$x^{(j)}$	n_{j1}	…	n_{ji}	…	n_{jc}	n_{j*}
…						…
$x^{(r)}$	n_{r1}	…	n_{ri}	…	n_{rc}	n_{r*}
Σ	n_{*1}	…	n_{*i}	…	n_{*c}	n

qualitative vs. qualitative

- Indices
  - $\chi^2 = n \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{\left(n_{ji} - \frac{n_{.i} * n_{j.}}{n} \right)^2}{n_{.i} * n_{j.}}$
    - Si X et Y indépendants, n_{ji} $n = n_{j*}$ n_{*i} et $\chi^2 = 0$
    - Plus χ^2 grand, plus la liaison entre X et Y est forte
  - $\Phi^2 = \frac{\chi^2}{n}$ indépendant de n
  - $T = \sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}$ Coefficient de Tschuprow (indép. de r et c)
  - $C = \sqrt{\frac{\Phi^2}{(min(r,c)-1)}}$ Coefficient de Cramer

Récapitulatif

quantitative vs. quantitative	quantitative vs. qualitative	qualitative vs. qualitative
nuage de points	boites parallèles	tableau de contingence
		
• r coefficient de corrélation linéaire	• $S_{Y\|X}$ rapport de corrélation	• χ^2

Vers le cas multidimensionnel: quantitatif

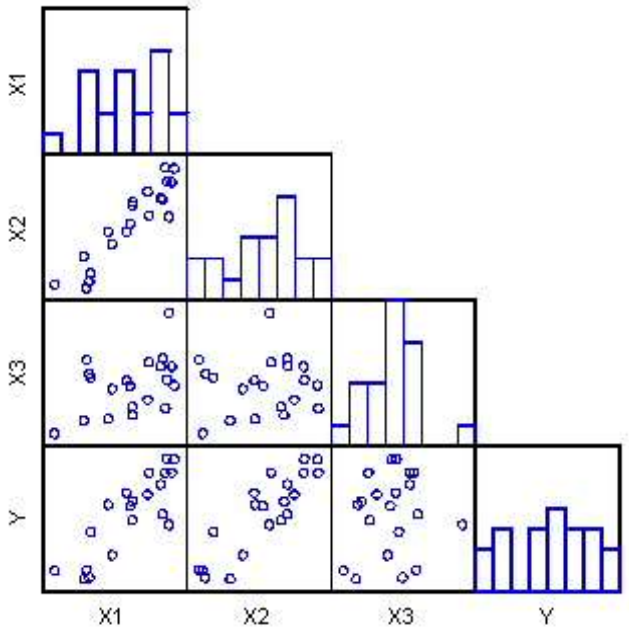
- $X_1,...,X_p$ p variables quantitatives observées sur le même échantillon

  - Matrice des variances-covariances $S_{(pxp)}$
    - S_{ii} = variance de X_i
    - S_{ij} = covariance de X_i et X_j

  - Matrice des corrélations $R_{(pxp)}$
    - $R_{ii} = 1$
    - R_{ij} = coefficient de corrélation linéaire entre X_i et X_j

Vers le cas multidimensionnel: quantitatif

- Tableau de nuages
  - Tab(i,i) = Histogramme(X_i)
  - Tab(i,j) = Nuage de points {X_i X_j}



Vers le cas multidimensionnel: qualitatif

- $X_1,...,X_p$ p variables qualitatives (de modalités respectives r_i) observées sur le même échantillon

  - Matrice des coefficients de Tschuprow (ou Cramer) $T_{(pxp)}$
    - $T_{ii} = 1$
    - T_{ij} = coefficient de Tschuprow entre X_i et X_j

  - Tableau de Burt
    - Tab(i,i) = Matrice diagonale (r_i x r_i) contenant les effectifs des r_i modalités de X_i
    - Tab(i,j) = Tableau de contingence entre X_i et X_j

Références

- Besse Ph.,
Statistiques uni et bidimensionnelles,
http://www.lsp.ups-tlse.fr/Besse/pub/sdm1.pdf

- Saporta G.,
Probabilités, analyse des données et statistique.
Technip, 1991.

- Wonnacott & Wonnacott,
Statistique : économie, gestion, science,
médecine, Economica, 1991.