Fouille de données

Cours 2 - Exploration des données : cas d'une et de deux dimensions

NGUYỄN Thị Minh Huyền ©2016

huyenntm@hus.edu.vn

Plan

Références

- Cours de Morin (IRISA, Rennes)- Chauchat (ERIC, Université Lyon 2) - cours donné à l'IFI en 2007
- Michel Tenenhaus, *Statistique: Méthodes pour décrire, expliquer et prévoir*, Dunod, 2007. Diapositives à l'adresse `https://studies2.hec.fr/jahia/Jahia/tenenhaus`
- Cours de Stéphane Tufféry `http://data.mining.free.fr/`
- J. Han and M. Kamber. `http://www.cs.illinois.edu/~hanj/bk3/`

Logiciels de Fouille de données

- Gratuits : Tanagra, Weka, R, etc.
- Payants : SAS, SPSS, S-Plus, etc.

Plan

Préparation de données

- Type de données
- Nettoyage du fichier (qualité des données)
- Distribution des variables
- Détection de valeurs aberrantes, extrêmes, rares, manquantes... et traitement
- Caractérisation des variables
- Création de nouvelles variables, transformation de variables

Exemple

- Exemple du Rola Cola de B.L. BOWERMAN / R.T. O'CONNELL (données fournies sur la page de Tenenhaus)
- Objectif: le département Marketing de Rola-Cola souhaite étudier les attitudes et les préférences des consommateurs envers Rola-Cola par rapport à Koca-Cola : pour cela, on réalise un test de goût avec les deux boissons avec des clients choisis au hasard.

Questions

1. Quelle boisson préférez-vous ?
   - Rola-Cola
   - Koka-Cola
2. Avez-vous déjà acheté Rola-Cola ?
   - Oui
   - Non
3. Entourez la réponse décrivant au mieux votre réaction à la phrase : J'aime mes boissons au Cola sucrées
   - D'accord
   - Je ne suis pas sûr
   - Pas d'accord
4. Combien de litres de boisson au Cola votre famille a-t-elle consommée au cours du mois dernier ?
5. Combien de paquets de chips avez-vous consommé le mois dernier ?

Données

- Fichier rola_cola.xls
- Echantillon : $n = 40$ personnes
- Codage :
  - Boisson préférée :
    1 = Rola-Cola   2 = Koka-Cola
  - Achat préalable :
    1 = oui   2 = non
  - Goût sucre :
    1 = oui   2 = indifférent   3 = non

Plan

Représentation de données

- Tableau
- Graphiques : diagramme circulaire (en secteurs), diagramme en bâtons, polygone de fréquence, histogramme, etc.

Etude d'une variable qualitative

- Etude d'une proportion
- Exemple : Boisson préférée entre Rola-Cola et Koca-Cola
  Feuille rola_cola.Proportion1

Etude d'une variable quantative (numérique)

- Une variable numérique X prend des valeurs réelles $x_1, \cdots, x_i, \cdots, x_N$ sur une population et $x_1, \cdots, x_i, \cdots, x_n$ sur un échantillon.
- Elle est résumée par des indicateurs statistiques :
  - Tendance centrale : moyenne, médiane, mode
  - Dispersion : étendues, écart-type, écart absolu moyen à la médiane...
  - Forme :
    Asymétrie (coefficient d'asymétrie : 0 - symétrique, > 0 - étalée à gauche , < 0 - étalée à droite)
    Aplatissement (coefficient d'aplatissement ou kurtosis : = 0 distribution normale, > 0 - concentration élevée, < 0 - concentration faible)

Etude d'une variable quantative (numérique)

- Une variable numérique X prend des valeurs réelles $x_1, \cdots, x_i, \cdots, x_N$ sur une population et $x_1, \cdots, x_i, \cdots, x_n$ sur un échantillon.
- Elle est résumée par des indicateurs statistiques :
  - Tendance centrale : moyenne, médiane, mode
  - Dispersion : étendues, écart-type, écart absolu moyen à la médiane...
  - Forme :
    Asymétrie (coefficient d'asymétrie : 0 - symétrique, > 0 - étalée à gauche , < 0 - étalée à droite)
    Aplatissement (coefficient d'aplatissement ou kurtosis : = 0 distribution normale, > 0 - concentration élevée, < 0 - concentration faible)

Etude d'une variable quantative (numérique)

- Une variable numérique X prend des valeurs réelles $x_1, \cdots, x_i, \cdots, x_N$ sur une population et $x_1, \cdots, x_i, \cdots, x_n$ sur un échantillon.
- Elle est résumée par des indicateurs statistiques :
  - Tendance centrale : moyenne, médiane, mode
  - Dispersion : étendues, écart-type, écart absolu moyen à la médiane...
  - Forme :
    Asymétrie (coefficient d'asymétrie : 0 - symétrique, > 0 - étalée à gauche , < 0 - étalée à droite)
    Aplatissement (coefficient d'aplatissement ou kurtosis : = 0 distribution normale, > 0 - concentration élevée, < 0 - concentration faible)

Tendance centrale : Mode, médiane

- Mode : valeur qui apparaît le plus fréquemment
- Médiane : M divise l'échantillon ordonné $x_1 \leq x_2 \leq \cdots \leq x_n$ en 2 parties égales
  - $n = 2k + 1$: $M = x_k$
  - $n = 2k$: $M = (x_k + x_{k+1})/2$

Tendance centrale et dispersion : Moyenne et écart-type

	Population	Echantillon
Effectif	N	n
Moyenne	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
Ecart-type	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

- $\overline{x}$ - estimation de μ,
- s^2 - estimation de σ^2.

Dispersion : Etendue, Quantiles

- Etendue = max - min
- Notion : Division de l'échantillon ordonné en *n* parties égales (quantiles)
  - $n = 4 \Rightarrow$ Quartiles Q_1, Q_2, Q_3 : charnières entre quatre parties.
    $Q_2 = M$, $Q_3 - Q_1$: étendue interquartile
  - $n = 10 \Rightarrow$ Déciles $D_1, \cdots, D_9$
    $D_9 - D_1$: étendue interdécile
  - $n = 100 \Rightarrow$ Centiles

Représentation graphique

- Proportion de la variable X : Diagrammes (en tuyaux d'orgue, en secteurs, en tige et feuilles), histogramme
- La dispersion de X est visualisée par la boîte-à-moustaches et l'histogramme.
  Boîte à moustaches : minimum, $[D_1]$, Q_1, médiane, Q_3, $[D_9]$, maximum
  $\Rightarrow$ aider à visualiser des valeurs extrêmes.

Exemple

Cas Rola-Cola

- Etude de la variable numérique : Comsommation de boisson au cola
- Statistiques et représentations graphiques
- Feuille rola_cola.Proportion2

Détection des observations atypiques (*Outliers*)

- La longueur de chaque moustache doit être inférieure à $1,5(Q_3 - Q_1)$.

Plan

Etude du lien entre deux variables

- 2 variables X et Y
  - X : variable explicative
  - Y : variable à expliquer
- 2 variables quantitatives : régression simple, corrélation simple
- 2 variables qualitatives : Test du khi-deux d'indépendance
- X quantitative, Y qualitative : régression logistique
- X qualitative, Y quantitative : analyse de la variance à un facteur

Deux variables quantitatives : nuage de points

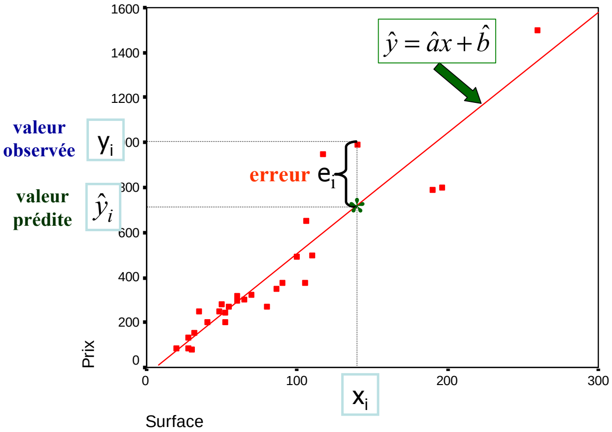
- Diagramme de dispersion
- Coefficient de corrélation
- Eventuellement, si cela a un sens, droite d'ajustement (des moindres carrés)

Données

- Y : variable à expliquer numérique (dépendante)
- X : variable explicative numérique ou binaire (indépendante)
- Tableau de données

	X	Y
1	x_1	y_1
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

La droite des moindres carrés



On cherche

$\boxed{\hat{a} \text{ et } \hat{b}}$

minimisant

$$\boxed{\sum_{i=1}^{n} e_i^2}$$

Coefficient de détermination R^2, coefficient de corrélation $Cor(X, Y)$

- Formule de décomposition :

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum e_i^2$$

- Coefficient de détermination :

$$R^2 = \frac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2}$$

- Coefficient de corrélation :

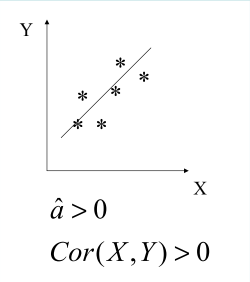
$$Cor(X, Y) = \text{sign}(\hat{a})\sqrt{R^2}$$

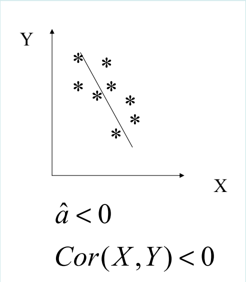
Corrélation entre deux variables : calcul direct de $Cor(X, Y)$

- Mesure la force et le sens de la liaison linéaire entre les deux variables numériques

$$Cor(X, Y) = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

- Toujours compris entre -1 et 1
- $Cor(X, Y) = 0$: X et Y non corrélées





Rappel sur les tests d'hypothèses

- Test d'hypothèses : raisonnement par l'absurde
- Hypothèse nulle H_0 : hypothèse inverse
- Objectif : calculer le degré de confiance en rejettant l'hypothèse nulle.

La corrélation $Cor(X, Y)$ est-elle significative au risque $\alpha = 0.05$?

- Notations
  - $\rho =$ corrélation au niveau de la population
  - $Cor(X, Y) =$ corrélation au niveau de l'échantillon
- Test :
  - H_0: $\rho = 0$
  - H_1: $\rho \neq 0$
- Règle de décision : On rejette H_0 au risque $\alpha = 0.05$ de se tromper si

$$|Cor(X, Y)| \geq \frac{2}{\sqrt{n}}$$

(Bonne approximation pour $n > 20$)

La corrélation $Cor(X, Y)$ est-elle significative au risque α ?

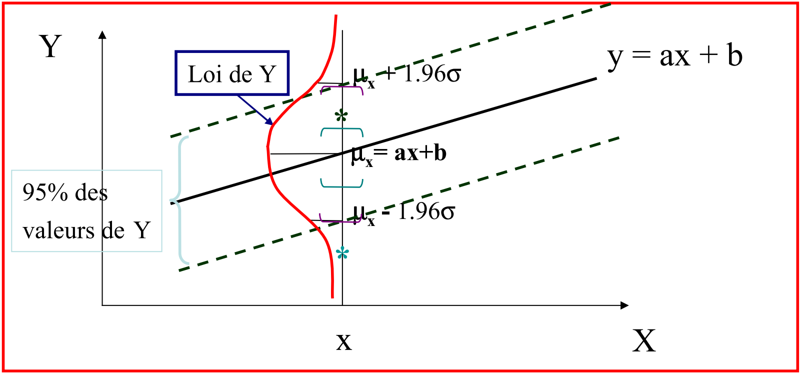
- Notations
  - $\rho =$ corrélation au niveau de la population
  - $Cor(X, Y) =$ corrélation au niveau de l'échantillon
- Test :
  - H_0: $\rho = 0$
  - H_1: $\rho \neq 0$
- Règle de décision : On rejette H_0 au risque α de se tromper si

$$|Cor(X, Y)| \geq \frac{t_{1-\alpha/2}(n-2)}{\sqrt{t^2_{1-\alpha/2}(n-2) + n - 2}}$$

- Plus petit α conduisant au rejet de H_0.

Modèle de la régression simple

- Modèle : $Y = aX + b + \epsilon$, avec $\epsilon \sim N(0, \sigma)$.



L'écart-type σ représente à peu près le quart de l'épaisseur du nuage.

Estimation de a, b et σ

- Estimation de a et b :
  - $\hat{a}$ = estimation de a
  - $\hat{b}$ = estimation de b
- Estimation de σ :
  - $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 =$ estimation de σ^2
  - $\hat{\sigma} = \sqrt{\hat{\sigma}^2} =$ estimation de σ

X et Y qualitatives

- On s'intéresse à l'indépendance entre les deux variables
- $\Rightarrow$ Test khi-deux (χ^2) de l'indépendance

Tableau de contingence

	1		j		p	
1	k_{11}		k_{1j}		k_{1p}	$k_{1\cdot}$
i	k_{i1}		k_{ij}		k_{ip}	$k_{i\cdot}$
n	k_{n1}		k_{nj}		k_{np}	$k_{n\cdot}$
	$k_{\cdot 1}$		$k_{\cdot j}$		$k_{\cdot p}$	$k = \sum k_{ij}$

Tableau de fréquences

$$f_{ij} = \frac{k_{ij}}{k}$$

	1		j		p	
1	f_{11}		f_{1j}		f_{1p}	$f_{1\cdot}$
i	f_{i1}		f_{ij}		f_{ip}	$f_{i\cdot}$
n	f_{n1}		f_{nj}		f_{np}	$f_{n\cdot}$
	$f_{\cdot 1}$		$f_{\cdot j}$		$f_{\cdot p}$	f

Lien entre deux variables

- Visualiser les associations entre les modalités des deux variables
- Tester l'indépendance entre les lignes et les colonnes
  - On observe k_{ij} ($k_{i.} = \sum_j k_{ij}$, $k_{.j} = \sum_i k_{ij}$, $k = \sum_{ij} k_{ij}$)
  - Si les variables sont indépendantes, alors $k_{ij}/k_{i.} = k_{.j}/k$ quel que soit i et $k_{ij}/k_{.j} = k_{i.}/k$ quel que soit j
  - Les $k_{ij}/k_{i.}$ sont appelés les profils lignes (il y en a autant que de lignes) et les $k_{ij}/k_{.j}$ les profils colonnes.
  - Sous l'hypothèse d'indépendance, $k_{ij} = k_{i.} * k_{.j}/k$

Comment étudier l'indépendance

- Examen des profils lignes ou colonnes
- Etude des d_{ij} = rapport observé/théorique $= k_{ij}/(k_{i.} * k_{.j}/k)$
- Statistique du χ^2 :

$$\chi^2 = \sum_{i,j} \frac{(k_{ij} - (k_{i.}k_{.j}/k))^2)}{k_{i.}k_{.j}/k}$$

A comparer à une valeur tabulée dans la table du khi-deux à $(n-1)(p-1)$ degrés de liberté.

Exemple

- Fichier Excel/Open Office Calc alcool.xls

X qualitative et Y quantitative

- Analyse de la variance (il faut que les écart-types soient les mêmes dans chaque groupe) - ANOVA
- De façon intuitive, si la variabilité entre groupes > la variabilité au sein d'un même groupe, on aura tendance à conclure que Y dépend des groupes. Si Y varie autant au sein d'un groupe qu'entre groupes, alors on aura tendance à conclure que X ne semble pas expliquer cette variabilité.
- L'ANOVA va permettre de fixer la limite (en fonction d'un risque α) à partir de laquelle on considère l'effet des groupes comme significatif.

ANOVA

- X définit k échantillons, dans chaque échantillon : n_i - effectif, $\bar{y}_i$ - moyenne, s_i - écart-type
- Global $n = \sum_{i=1}^{k} n_i$; moyenne générale $\bar{y} = \sum_{i=1}^{k} n_i \bar{y}_i / n$
- Y_i : variable Y sur la population i, chaque Y_i suit une loi normale $N(\mu_i, \sigma)$
- Somme des carrés intra-groupe : $\text{ssw} = \sum_{i=1}^{k} (n_i - 1) s_i^2 / (n - k)$
- Somme des carrés inter-groupes : $\text{ssb} = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 / (k - 1)$
- Rapport de corrélation : $\eta^2 = \text{ssb}/\text{ssw} = F$
- F-Test : $F \geq F_{1-\alpha}(k-1, n-k)$
- Exemple : fichier MS Excel/Open Office Calc iris.xls

X quantitative et Y qualitative

Régression logistique

- Valeurs de la variable à prédire Y sont binaires (0 ou 1)
- Au lieu de prédire la valeur de Y, on prédit $P(Y = 0|X)$ ou $P(Y = 1|X)$.
- Les probabilités décrivent une sigmoïde (courbe en forme de S) entre 0 et 1
- $$P(Y = 1|x_1, x_2, \cdots, x_k) = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}$$
- β_i à estimer par des programmes (utilisant des méthodes comme MLE - Maximum Likelihood Estimate ou Newton-Raphson)
- $\beta_i = 0$: pas d'effet sur la chance de succès, $\beta_i > 0$: augmente la chance, $\beta_i < 0$: décroît la chance

Résumé - Objectifs

- Préparation et exploration des données
- Nettoyage des données
  - Valeurs extrêmes : transformation, élimination ?
  - Valeurs manquantes : élimination, remplacement (valeur moyenne, régression) ?
- Etape très importante (conditionne la fiabilité de la suite).
- Ce cours : cas d'une ou deux variables
- Cas de plus de 2 variables : cours suivant.

Travail à faire

- Travail en groupe
- Exploration d'un fichier de données (au choix) avec Tanagra