TD 2 — Cours

Pierre Beaucoral

2025-09-15

Introduction

Ce document reprend **strictement le contenu** du fichier de slides fourni (*td2-slides.qmd*), réorganisé en **format cours** (HTML) avec une courte phrase narrative au début de chaque section pour guider la lecture. Aucune matière externe n'a été ajoutée. ## Gestion de la base de données

Ces diapositives introduisent les statistiques descriptives : elles servent à résumer et visualiser les variables avant toute modélisation (tendance centrale, dispersion, forme, comparaisons par modalité).

Gestion de la base de données

Commandes de base

Ces commandes se tapent dans la **fenêtre de commande** d'EViews.

Commande	Description	Exemple
`group`	Créer un **groupe** de variables	`group nom x y`
`scalar`	Créer un **scalaire** et faire des calculs	`scalar k = 3*6`
`matrix`	Créer une **matrice** et faire des calculs (matriciels)	*(cf. Help)*
`genr`	**Générer** une variable	*(cf. plus bas)*
`rename`	**Renommer** une variable	`rename x y` *(renomme x en y)*
`delete`	**Effacer** un ou plusieurs objets	`delete x y`

Commande	Description	Exemple
`smpl`	**Sélectionner** un **sous-échantillon**	`smpl if x<10`

Type de variables

Vous êtes souvent amenés à **créer** de nouvelles variables ou à en changer l'**échelle**.
Il existe **deux types** de variables :

1. **Variable continue**
   prend n'importe quelle valeur sur un intervalle donné.
2. **Variable discrète**
   ne prend qu'un nombre ***fini*** *de valeurs.*

> **Note**
>
> **Nota** : les variables **binaires** (0/1) sont un cas particulier.

Création de variables continues

Commande générale :

- `genr nouveau_nom = opération`
  Ex. `genr lnx = log(x)`

> **Warning**
>
> **Attention** : saisir la ligne de commande dans la **fenêtre de commande**.

Opérateurs utiles (EViews)

Formule mathématique	EViews
$x + a$	`x+a`
$x - a$	`x-a`
$x \cdot a$	`x*a`
x/a	`x/a`
x^a	`x^a`
$\ln(x)$	`log(x)`
e^x	`exp(x)`

Création de variables muettes (dummies)

Une variable muette est **discrète 0/1** (binaire).

Exemples : - (=1) si l'individu est une **femme**, (0) sinon. - (=1) si le pays est **OCDE**, (0) sinon.

Deux **méthodes** :

Méthode 1 — rapide (condition)

- `genr nouveau = x > A`
  Ex. **jeune** vaut 1 si âge ≤ 25, 0 sinon :
  `genr jeune = age <= 25`

Méthode 2 — en plusieurs étapes

Objectif : **riche** vaut 1 si `pibtete > 10000`, 0 sinon.

1. `genr riche = 0`
2. `smpl if pibtete > 10000`
3. `genr riche = 1`
4. `smpl @all`

Création de variables discrètes

Une variable **discrète** prend un nombre **limité** de valeurs (0, 1, ..., n).
- Peut venir d'un **classement** (ex. classes de **revenus**).
- Peut coder un **choix** limité (pays d'immigration, parti politique, notes...).

> **i** Note
>
> **Remarque** : une **muette** est un cas particulier de **discrète**.

Deux **méthodes** :

Discrètes — Méthode 1 (somme de dummies)

Exemple : classes d'**âge** (

$$classes = \begin{cases} 0 & si \quad age \leq 25 \\ 1 & si \quad 25 < age \leq 35 \\ 2 & si \quad 35 < age \leq 45 \\ 3 & si \quad 45 < age \end{cases}$$

)

Dans la **fenêtre de commande** :

```
genr dummy1 = age > 25
genr dummy2 = age > 35
genr dummy3 = age > 45
genr classes = dummy1 + dummy2 + dummy3
```

Discrètes — Méthode 2 (par étapes)

On **réplique** la méthode 2 des muettes :

1. `genr classes = 0`
2. `smpl if condition1` → `genr classes = 1` → `smpl @all`
3. `smpl if condition2` → `genr classes = 2` → `smpl @all`
4. etc.

Exploration de la base de données

Principe

Le but des **statistiques descriptives** est de **décrire** les variables.
Étape **cruciale** pour :
- connaître sa base,
- avoir une première idée des **relations** existantes.

Étude d'une variable	Étude d'une relation entre variables
Tableaux : statistiques descriptives **Figures** : histogramme, boîte à moustache, évolution	**Coefficients** de corrélation **Nuage de points**, droite de régression

---

Statistiques descriptives — une variable

Ouvrir la fenêtre de la série (double-clic).

- **Tableau** des principales statistiques :
  `View → Descriptive statistics & Tests → Stats Table`
- **Graphiques** : `View → Graph`
  - **Histogramme** : distribution
  - **Line** : courbe temporelle
  - **Boxplot** : boîte à moustache

Comparer par modalité (une variable)

Test d'égalité de moyennes :
`View → Descriptive statistics & Tests → Stats by classification`
- Choisir la **modalité** via *Series/Group for classify*.

Graphiques par modalité :
- Option **Categorical graph** dans *Graph type*.
- Renseigner la modalité dans *factors — series defining categories*.

Statistiques descriptives — plusieurs variables

Ouvrir les séries ensemble : sélectionner les variables → `Open → as Group`.

- **Coefficients de corrélation** :
  1. `View → Covariance analysis`
  2. Dans **Statistics**, choisir **Correlation**
- **Nuage de points** :
  `View → Graph → Scatter`
  - Utile : **Fit Line → Regression line** (droite de régression)

Graphiques (plusieurs variables)

Explorer les différents **graphiques** et **choisir** celui qui illustre le mieux votre propos.

- Pour **modifier** le graphique : bouton **Options** (fenêtre du graphique)
- Pour **restreindre** à un sous-échantillon : onglet **Sample**

Enregistrer et extraire les objets créés

Pour **enregistrer** les objets : **Freeze** et **nommer** (Name).

- **Tableaux** : le plus simple → **Copy** (Ctrl+C) et **coller** dans **Excel**.
- **Graphiques** :
  `Proc → Copy to Clipboard` (ou Ctrl+C)
  ou `Object → View Options → Copy to Clipboard`
  puis coller dans un document **Word (.doc)**.

Questions – Réponses (TD2)

Question : Importez la base de données TD2.xls.

Afficher la réponse

On créer le fichier workfile et on fait : **file → workfile** et ensuite on fait **file → Import → import from file**.

Question : Générez le nombre d'accidents (mortels et non mortels) pour chaque compagnie.

Afficher la réponse

```
genr Accidents = fatal + non_fatal
```

Question : Rapportez ce nombre d'accidents au nombre de passagers transportés : Quelle est l'utilité de cette transformation ?

Afficher la réponse

Commande :

```
genr Acc_pass = Accidents / passagers
```

Utilité : Rapporter le nombre d'accidents au nombre de passagers permet d'évaluer le risque d'accident par passager, offrant un indicateur plus précis de la sécurité des compagnies, indépendamment de leur taille. Autrement dit, cela permet d'évaluer de manière précise la probabilité d'accident par rapport au nombre total de passagers.

Question : Créez une variable prenant la valeur de 1 si la compagnie a connu au moins un accident au cours des 15 dernières années.

Afficher la réponse

```
genr Dummy_acc = accidents >= 1
```

Question : Même question en distinguant entre accidents mortels et non mortels.

Afficher la réponse

```
genr Dummy_fatal = fatal >= 1 genr Dummy_non_fatal = non_fatal >= 1
```

Question : Ouvrez la variable dépendante (passagers) et étudiez sa distribution : Que pouvez-vous en conclure (concentration, points aberrants) ?

Afficher la réponse

Pour voir la distribution, on utilise un **histogramme**.
Dans EViews : *on clique sur la variable* **passagers** → **View** → **Graph** → **Distribution** → **OK**.
Pour copier-coller : **Proc** → **Copy to Clipboard (ou Ctrl+C)** ou **Object** → **View Options** → **Copy to Clipboard**, puis coller dans un document Word (.doc).

L'histogramme montre une **distribution asymétrique à droite**, indiquant que la plupart des valeurs sont concentrées à gauche, tandis qu'il y a quelques valeurs élevées moins fréquentes à droite. Cet histogramme montre une distribution très asymétrique avec une concentration élevée des données à gauche, ce qui indique que la majorité des valeurs observées sont faibles. À l'inverse, on observe que les valeurs plus élevées sont rares, avec quelques points dispersés à droite (points **aberrants**).

Regardez si la distribution de la variable *passagers* diffère :

- si le pays a connu au moins un accident
- si le pays a connu au moins un accident **mortel**.

Afficher la réponse

Dans EViews :

1. **Ouvrir la série** `passagers` → `View → Descriptive statistics & Tests → Stats by classification`.
2. Dans *Series/Group for classify*, sélectionner la variable indiquant :
   - (i) s'il y a eu **au moins un accident**,
   - (ii) s'il y a eu **au moins un accident mortel**.
3. Valider pour obtenir les tableaux de statistiques et, si souhaité, les **graphes par modalité** (*Graph type* → *Categorical graph*).

Interprétation : Comparer les statistiques (moyenne, médiane, etc.) et les graphiques permet de voir si le nombre de passagers transportés est distribué différemment selon qu'il y a eu un accident ou un accident mortel.

On observe généralement une **différence nette des moyennes** : les compagnies ayant connu un (ou un accident mortel) présentent en moyenne un **volume de passagers plus élevé**, ce qui suggère qu'elles sont plus grandes et donc exposées à un risque absolu d'accident plus important.

Question : Regardez la corrélation entre le nombre de passagers transportés et l'âge de la compagnie à l'aide des coefficients de corrélation et d'un nuage de points : Ces deux variables sont-elles fortement liées ? Pourquoi ?

Afficher la réponse

Création de l'âge : `genr age = 2013 - annee`

Corrélation : sélectionner les deux variables en **Group** → **View** → **Covariance analysis** → *Statistics* = **Correlation**.
On a un coefficient de corrélation de **-0,1709** qui indique une **faible corrélation positive** entre les deux variables étudiées (relation faible, peut ne pas être significative).

Nuage de points : sélectionner les deux variables → **View** → **Graph** → **Scatter**. Les points montent de gauche à droite (tendance conjointe) et des points qui s'écartent du nuage principal peuvent indiquer des **valeurs aberrantes**.
