11. Révision

Sam Gyetvay

ECO 2273 – Économetrie I

3 décembre 2025

Séance de révision d'aujourd'hui

L'examen final couvrira toute la matière du cours. Aujourd'hui, nous allons réviser :

- Données et statistiques descriptives
- Théorie des probabilités
- Échantillonnage et estimation
- Intervalles de confiance
- Tests d'hypothèses
- Régression linéaire

Ensuite, nous résoudrons quelques questions pratiques (disponibles sur le site web du cours)

Données

Nom	Sexe	Âge	Lieu de naissance	Éducation
Sam	M	34	Montréal, Canada	PhD
Alex	M	29	Toronto, Canada	BA
Marie	F	31	Québec, Canada	MA

Chaque ligne = une observation (personne)
Chaque colonne = une variable (Nom, Sexe, Âge, Lieu de naissance, Éducation)

La première ligne n'est pas une observation, elle donne simplement les noms des variables

Statistiques descriptives

Certaines statistiques descriptives nous renseignent sur la **tendance centrale** d'une variable :

- **Moyenne** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ — la moyenne arithmétique
- **Médiane** — la valeur du milieu lorsque triées

D'autres statistiques descriptives nous renseignent sur la **dispersion** d'une variable :

- **Variance** $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — écart quadratique moyen par rapport à la moyenne
- **Écart-type** $s = \sqrt{s^2}$ — racine carrée de la variance

Covariance et corrélation

Pour mesurer la relation entre deux variables X et Y :

Covariance :

$$Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Corrélation :

$$\rho_{xy} = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

- La corrélation est toujours entre -1 et +1 : $-1 \leq \rho_{xy} \leq 1$
- Corrélation positive : X et Y « évoluent ensemble »
- Corrélation négative : X et Y « évoluent dans des directions opposées »

Théorie des probabilités : révision rapide

Événements et probabilités

- Un **événement** est quelque chose qui peut se produire (par ex., obtenir un 6)
- La **probabilité** $P(A)$ mesure la probabilité qu'un événement A se produise
- $0 \leq P(A) \leq 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B) = P(A) \cdot P(B|A)$
- Les événements A et B sont **indépendants** si $P(A \cap B) = P(A) \cdot P(B)$

Variables aléatoires

- Une **variable aléatoire** X associe un nombre à chaque événement
- Peut être **discrète** (par ex., nombre de faces dans 10 lancers de pièce) ou **continue** (par ex., taille en cm)

Variables aléatoires et distributions

Espérance et variance

- Espérance $E[X]$ = valeur moyenne de X
- Variance $Var(X)$ = dispersion de X
- Propriétés : $E[aX + b] = aE[X] + b$, $Var(aX + b) = a^2 Var(X)$

Distributions célèbres

- **Binomiale** — nombre de succès dans n essais indépendants
- **Poisson** — nombre d'événements dans une période de temps fixe
- **Uniforme** — toutes les valeurs dans une plage également probables
- **Normale** — courbe en cloche, la plupart des valeurs près de la moyenne

La distribution normale centrée réduite

La distribution **normale centrée réduite** $Z \sim N(0, 1)$ a une moyenne 0 et une variance 1

Toute variable aléatoire normale peut être **normalisée** pour devenir normale centrée réduite :

$$\text{Si } X \sim N(\mu, \sigma^2), \text{ alors } Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Quantiles clés : $z_{0.975} = 1.96 \approx 2$, $z_{0.95} = 1.645$

- $P(Z \leq z_{0.975}) = 0.975$
- $P(Z \leq z_{0.95}) = 0.95$

Population vs Échantillon

Population = tous les individus que nous voulons étudier

- Si nous menions un recensement, nous pourrions calculer la vraie moyenne $E[X] = \mu$ et la variance $Var(X) = \sigma^2$ dans la population

Puisque nous ne pouvons pas observer toute la population, nous devons prendre des échantillons

- Nous observons n individus de la population
- Calculer la moyenne d'échantillon $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Calculer la variance d'échantillon $s^2 = \frac{1}{n-1}(X_i - \bar{X})^2$

En statistique, les vraies valeurs μ, σ^2 sont appelées paramètres de population. Ce sont des nombres fixes inconnus. Nous utilisons $\bar{X}$ et s^2 pour estimer la moyenne μ et la variance σ^2 de la population. Puisque $\bar{X}$ et s^2 dépendent de l'échantillon que nous tirons, nos estimations sont bruitées et incertaines

Échantillonnage aléatoire et absence de biais

L'**échantillonnage aléatoire** signifie que chaque individu a une chance égale d'être sélectionné

Lorsque nous utilisons un échantillonnage aléatoire, la moyenne d'échantillon $\bar{X}$ est un **estimateur sans biais** de la moyenne de population μ :

$$E[\bar{X}] = \mu$$

En moyenne (sur tous les échantillons possibles) $\bar{X}$ ne sur-estime ni ne sous-estime μ

Un échantillon particulier pourrait donner $\bar{X} > \mu$ ou $\bar{X} < \mu$

Mais si nous échantillonnions plusieurs fois et faisions la moyenne de toutes les valeurs $\bar{X}$, nous obtiendrions μ

Théorème central limite (TCL)

Le **Théorème central limite** stipule que pour n grand :

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- La moyenne d'échantillon est approximativement normale (même si X_i ne l'est pas !)
- La variance diminue avec n : $Var(\bar{X}) = \frac{\sigma^2}{n}$
- Pour des échantillons avec n plus grand, $\bar{X}$ sera plus proche de la vraie valeur μ en moyenne
- Fonctionne pour toute distribution de X_i : même si X_i n'est pas normale
- La variance ne dépend que de σ^2 et n !

Erreur-type vs écart-type

Ce sont deux concepts différents mais liés que les étudiants confondent souvent :

Écart-type σ mesure la dispersion de la variable X

- Combien les observations individuelles X_i varient-elles autour de la moyenne de population μ ?
- $\sigma = \sqrt{Var(X)}$

Erreur-type $\frac{\sigma}{\sqrt{n}}$ mesure la dispersion de la moyenne d'échantillon $\bar{X}_n$

- Combien notre estimation $\bar{X}_n$ varie-t-elle d'un échantillon à l'autre ?
- $\frac{\sigma}{\sqrt{n}} = \sqrt{Var(\bar{X}_n)}$
- À mesure que n augmente, l'erreur-type diminue : plus de données = estimations plus précises

L'erreur-type nous indique à quel point nous devrions être confiants dans notre estimation $\bar{X}_n$

La distribution de Student

Lorsque nous ne connaissons pas σ et devons l'estimer avec notre échantillon s, nous utilisons la **distribution de Student** au lieu de la distribution normale pour sélectionner les valeurs critiques

La distribution t a un paramètre : **degrés de liberté** (dl)

- Pour tester μ : dl = $n - 1$
- Pour tester β_1 en régression : dl = $n - 2$

Lorsque n est grand, la distribution de Student est presque identique à la normale centrée réduite. Par conséquent, nous pouvons utiliser les approximations $t_{0.975,n-1} \approx z_{0.975} = 1.96 \approx 2$ et $t_{0.95,n-1} \approx z_{0.95} = 1.645$

Lorsque n est petit, l'approximation normale est mauvaise, et nous devons utiliser $t_{0.975,n-1}$ et $t_{0.95,n-1}$. Dans le test, celles-ci seront données (voir le test pratique pour un exemple)

Intervalle de confiance à 95% pour μ

Un intervalle de confiance à 95% pour la moyenne de population μ est :

$$\bar{X} \pm t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}}$$

Pour n grand, $t_{0.975,n-1} \approx 2$, donc :

$$\bar{X} \pm 2 \cdot \frac{s}{\sqrt{n}}$$

Rejeter $H_0 : \mu = \mu_0$ au niveau 5% si μ_0 n'est pas contenu dans l'intervalle de confiance à 95%

Intervalles de confiance pour les proportions

Lorsque X_i est binaire $X_i \in \{0, 1\}$, sa moyenne $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ est une proportion $\hat{p}$. Dans ce cas, l'intervalle de confiance à 95% est

$$\hat{p} \pm t_{0.975,n-1} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

où :

- $\hat{p}$ est la proportion d'échantillon
- $t_{0.975,n-1} \approx 2$ pour IC à 95%

Erreurs de type I et de type II

	H_0 est vraie	**H_0 est fausse**
Rejeter H_0	Erreur de type I (Faux positif)	Correct (Vrai positif)
Ne pas rejeter H_0	Correct (Vrai négatif)	Erreur de type II (Faux négatif)

Erreur de type I : Rejeter une hypothèse nulle vraie

- Aussi appelée « faux positif »
- Exemple : condamner une personne innocente

Erreur de type II : Ne pas rejeter une hypothèse nulle fausse

- Aussi appelée « faux négatif »
- Exemple : ne pas condamner un criminel

Test d'hypothèse bilatéral : La recette

Pour effectuer un test **bilatéral** de $H_0 : \mu = \mu_0$ au niveau de signification de 5% :

Étape 1 : Calculer la statistique de test

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

où s est l'écart-type de l'échantillon.

Étape 2 : Comparer à la valeur critique

- Si $|t| > t_{0.975,n-1} \approx 2$, **rejeter** H_0
- Si $|t| \leq t_{0.975,n-1} \approx 2$, **ne pas rejeter** H_0

Note : Pour n grand, nous pouvons utiliser $z_{0.975} = 1.96 \approx 2$

Tests d'hypothèses unilatéraux

Pour les tests **unilatéraux**, l'hypothèse nulle prend la forme :

$$H_0 : \mu \leq \mu_0$$

- Rejeter si $t > t_{0.95,n-1}$
- Valeur critique : $t_{0.95,n-1} \approx 1.645$ (pour n grand)

$$H_0 : \mu \geq \mu_0$$

- Rejeter si $t < -t_{0.95,n-1}$
- Valeur critique : $-t_{0.95,n-1} \approx -1.645$ (pour n grand)

Pour un test unilatéral, utiliser $t_{0.95}$ au lieu de $t_{0.975}$

Test z vs test t : Quand utiliser lequel ?

Test z : Utiliser lorsque σ (écart-type de population) est connu

- Statistique de test : $z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
- Utiliser $z_{0.975} = 1.96$, $z_{0.95} = 1.645$ comme valeurs critiques

Test t : Utiliser lorsque σ est inconnu

- Statistique de test : $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
- $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- Utiliser $t_{0.975,n-1}$, $t_{0.95,n-1}$ comme valeurs critiques
- Pour n grand : $t_{0.975,n-1} \approx 1.96$, $t_{0.95,n-1} \approx 1.645$

Dans l'examen, utiliser les approximations pour grand échantillon ($t_{0.975,n-1} \approx 1.96$, $t_{0.95,n-1} \approx 1.645$) sauf si n est très petit ($n < 10$). Lorsque n est petit, je vous donnerai des valeurs critiques, vous devez choisir la bonne.

Test d'égalité des moyennes (deux échantillons)

Pour tester si deux groupes ont des moyennes égales ($H_0 : \mu_1 = \mu_2$) :

Statistique de test :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

où :

- $\bar{X}_1, \bar{X}_2$ = moyennes d'échantillon
- s_1, s_2 = écarts-types d'échantillon
- n_1, n_2 = tailles d'échantillon

Règle de décision (niveau 5%) :

- Si $|t| > 2$, rejeter H_0 (les moyennes sont différentes)
- Si $|t| \leq 2$, ne pas rejeter H_0

Tables d'équilibre dans les expériences randomisées

Une **table d'équilibre** vérifie si la randomisation a réussi en comparant les caractéristiques de base entre les groupes de traitement et de contrôle.

Caractéristique	Traitement	Contrôle	Diff	valeur-p
Âge (années)	32.5	32.8	-0.3	0.65
Éducation (années)	12.2	11.8	0.4	0.52
Femme (proportion)	0.48	0.52	-0.04	0.42
Salaire antérieur ($/h)	18.5	22.1	-3.6	0.92

La valeur-p correspond à l'hypothèse nulle $H_0 : \mu_T = \mu_C$. Parfois, les tables d'équilibre incluent les erreurs-types, les statistiques-t, ou les intervalles de confiance à 95% au lieu des valeurs-p

Régression linéaire simple

Une régression linéaire simple est un modèle de la forme

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$$

- $i \in \{1, \ldots, n\}$ indexe les observations
- y_i est la **variable dépendante**
- x_i est la **variable indépendante**
- ε_i est l'**erreur**
- β_0 et β_1 sont les **coefficients**
    - β_0 est l'**ordonnée à l'origine** ou la **constante**
    - β_1 est la **pente**

« Simple » fait référence au fait qu'il n'y a qu'une seule variable indépendante

Estimation des coefficients par moindres carrés

Étant donné les données $(y_i, x_i)_{i=1}^n$, les estimations par **moindres carrés ordinaires (MCO)** sont :

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

La droite de régression passe *toujours* par le point $(\bar{x}, \bar{y})$:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Relation entre $\hat{\beta}_1$, corrélation et écarts-types

Le coefficient de pente MCO $\hat{\beta}_1$ est lié au coefficient de corrélation ρ_{xy} par :

$$\hat{\beta}_1 = \rho_{xy} \cdot \frac{\sigma_y}{\sigma_x}$$

où σ_x et σ_y sont les écarts-types de x et y.

En réarrangeant :

$$\rho_{xy} = \hat{\beta}_1 \cdot \frac{\sigma_x}{\sigma_y}$$

R^2 : Mesurer la qualité d'ajustement

Le R^2 (R carré) mesure la fraction de la variance de y expliquée par x :

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- $R^2 \in [0, 1]$
- $R^2 = 0$: x n'explique aucune variation de y
- $R^2 = 1$: x explique toute la variation de y
- Un R^2 plus élevé signifie un meilleur ajustement (plus de pouvoir explicatif)

Régression de traitement binaire

Lorsque la variable indépendante est une variable binaire (par ex., $D_i \in \{0, 1\}$ pour traitement/contrôle) :

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

Alors :

- $\hat{\beta}_0$ = moyenne de y pour le groupe de contrôle ($D = 0$)
- $\hat{\beta}_0 + \hat{\beta}_1$ = moyenne de y pour le groupe de traitement ($D = 1$)
- $\hat{\beta}_1$ = différence de moyennes entre traitement et contrôle

Formule clé :

$$\hat{\beta}_1 = E[y|D = 1] - E[y|D = 0] = \bar{y}_{\text{traitement}} - \bar{y}_{\text{contrôle}}$$

Loi des espérances itérées (LEI)

La **Loi des espérances itérées** stipule :

$$E[E[Y|X]] = E[Y]$$

En mots : L'espérance d'une espérance conditionnelle est égale à l'espérance inconditionnelle.

Lorsque X est discrète : la moyenne globale est une moyenne pondérée des moyennes de sous-groupes :

$$E[Y] = \sum_{x} E[Y|X = x] \cdot P(X = x)$$

Erreur-type de $\hat{\beta}_1$

L'erreur-type de $\hat{\beta}_1$ est

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{1}{n-2} \cdot \frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i (x_i - \bar{x})^2}}$$

où $\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ sont les résidus

L'estimation $\hat{\beta}_1$ sera moins précise lorsque la variance des résidus est plus élevée par rapport à la variance de X

$$\frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i (x_i - \bar{x})^2} = \frac{\frac{1}{n} \sum_i (\hat{\varepsilon}_i - 0)^2}{\frac{1}{n} \sum_i (x_i - \bar{x})^2} = \frac{Var(\hat{\varepsilon})}{Var(x)}$$

Test d'hypothèse de $\beta_1 = 0$ au niveau 5%

Nous voulons tester

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_A : \beta_1 \neq 0.$$

Statistique de test :

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}.$$

Règle de décision (niveau 5%, bilatéral) :

$$\text{Rejeter } H_0 \quad \text{si} \quad |t| > t_{n-2,\,0.975}.$$

La valeur critique $t_{n-2,\,0.975}$ est le 97.5e percentile de la distribution t avec $n-2$ degrés de liberté. Lorsque n est grand, nous pouvons utiliser l'approximation $t_{n-2,\,0.975} \approx 2$

Intervalle de confiance à 95% pour β_1

Un intervalle de confiance à 95% pour β_1 est :

$$\hat{\beta}_1 \pm t_{0.975,n-2} \cdot SE(\hat{\beta}_1)$$

Pour n grand, $t_{0.975,n-2} \approx 2$, donc :

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

Rejeter $H_0 : \beta_1 = 0$ au niveau de signification de 5% si l'intervalle de confiance à 95% ne contient pas zéro

Lire la sortie de régression STATA

```
      Source |       SS           df       MS          Number of obs   =     250
-------------+----------------------------------       F(1, 248)       = 112.50
       Model |   2250.00            1   2250.00        Prob > F        =  0.000
    Residual |   4960.00          248     20.00        R-squared       = 0.3125
-------------+----------------------------------       Adj R-squared   = 0.3097
       Total |   7210.00          249     28.96        Root MSE        = 4.4721

------------------------------------------------------------------------------
        wage | Coefficient  Std err       t     P>|t|      [95% conf. interval]
-------------+----------------------------------------------------------------
   education |     3.00       0.40       7.50   0.000      [ 2.21,    3.79 ]
       _cons |    10.00       2.50       4.00   0.000      [ 5.08,   14.92 ]
------------------------------------------------------------------------------
```

- Coefficients : $\hat{\beta}_1 = 3.00$, $\hat{\beta}_0 = 10.00$
- Erreurs-types : $\hat{\sigma}_{\hat{\beta}_1} = 0.40$, $\hat{\sigma}_{\hat{\beta}_0} = 2.50$
- Statistique-t, valeur-p pour $H_0 : \beta_k = 0$
- $R^2 = 0.3125$