

---

# LNG-1100 : Méthodes expérimentales et analyse de données

ANOVA → régression linéaire

Guilherme D. Garcia

[fr.gdgarcia.ca](http://fr.gdgarcia.ca) ↗

6–7



# Plan de la séance

Dans RStudio aujourd'hui

1. ANOVA
2. Régression linéaire
3. Extra : l'erreur standard et le coefficient de détermination

☞ Vous devez choisir votre équipe pour les problèmes 1 et 2 (28 octobre)



# PARTIE 1



### Pratique

- Vous voyez  $F(4, 203)$  dans un article scientifique. Combien de groupes et d'observations y a-t-il dans l'étude?
- Pourquoi ajustons nous les valeurs  $p$  dans un test *post hoc*?
- L'effet d'une variable est-il présent dans le résultat d'une ANOVA? Pouvons-nous le déduire?



# ANOVA

## Révision

- **One-way** ANOVA →  $y \sim x$  p. ex., note en fonction de ville
- **Two-way** ANOVA →  $y \sim x + w$  p. ex., note en fonction de ville et langue maternelle
- **ANCOVA** →  $y \sim x + z$  p. ex., note en fonction de ville et nombre d'heures d'étude<sup>1</sup>

- ☞ On a vu un exemple d'une one-way ANOVA seulement, mais la logique est la même
  - Et vous pouvez utiliser toujours la même fonction : `aov( ... )`
  - L'ANCOVA = plus général : en pratique, on a souvent des variables catégoriques et continues

<sup>1</sup>Donc, une variable continue ( $z$ ).



# ANOVA → régression

- ANOVA : type spécial de **régression** — méthode très générale d'analyse de données
  - ☞ Dans le cours, on utilisera la régression traditionnelle parce que :
    1. elle est plus **puissante** et plus **complète**
    2. ses résultats mettent l'accent sur l'**effet**, et non pas sur les valeur *p*
    3. la méthode est facilement généralisable
    4. on veut **prédire** des données
- **Mas c'est quoi une « régression »?**



# Régression

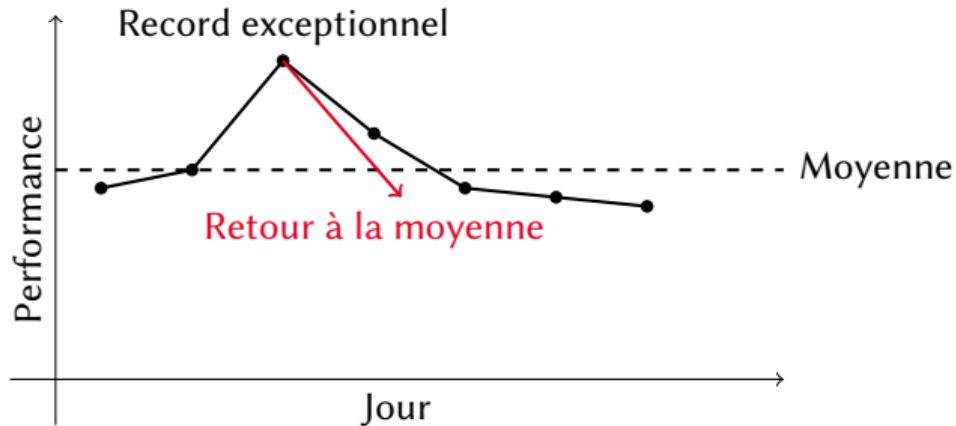
## Un processus naturel

- Un sprinter de haut niveau qui réalise un temps **exceptionnel** lors d'une course
  - Lors des courses suivantes, son temps revient souvent à sa moyenne habituelle
- ☞ **Pourquoi ?** La performance exceptionnelle est due en partie à des facteurs aléatoires :
- Conditions météorologiques favorables
  - Forme physique particulièrement bonne ce jour-là
  - Un simple coup de chance
  - ...
- Lors des prochaines courses, ces éléments fluctuants ne joueront pas forcément en sa faveur, et il reviendra à un niveau **plus représentatif** de ses capacités
- ☞ **Bref :** les performances extrêmes ont tendance à être suivies d'un retour à la normale



# Régression vers la moyenne

Un processus naturel : figure à titre d'illustration

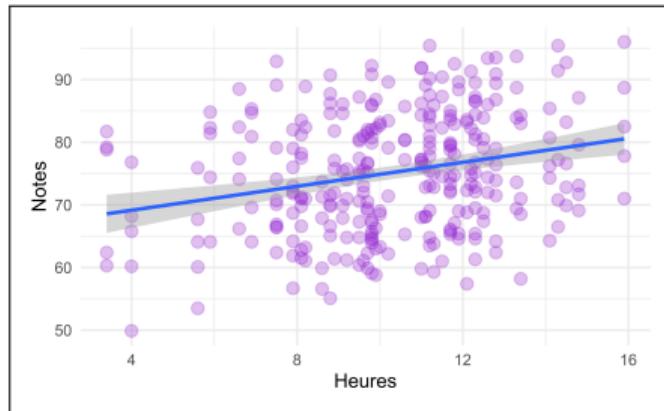


- ☞ Plus une performance est exceptionnelle, plus il est probable que les prochaines performances seront moins exceptionnelles (on a un maximum, après tout!)



# Régression linéaire

- Voici une figure qui contient une régression linéaire (`feedbackData.csv`) :



- Ici, on examine l'effet des **heures** (**variable continue**) d'étude sur la note
- ☞ « Quelle sera la différence dans la note pour chaque heure d'étude additionnelle? »



# Régression linéaire

- Quelle est la meilleure droite pour les données?
- « Meilleure » → celle qui minimise les distances entre les points et elle-même

☞ On utilise deux paramètres pour définir la droite :

l'**intercept** =  $\beta_0$  (l'ordonnée à l'origine)

le **slope** =  $\beta_1$  (la pente)

- On peut visualiser ces paramètres [ici](#) (Trend line)



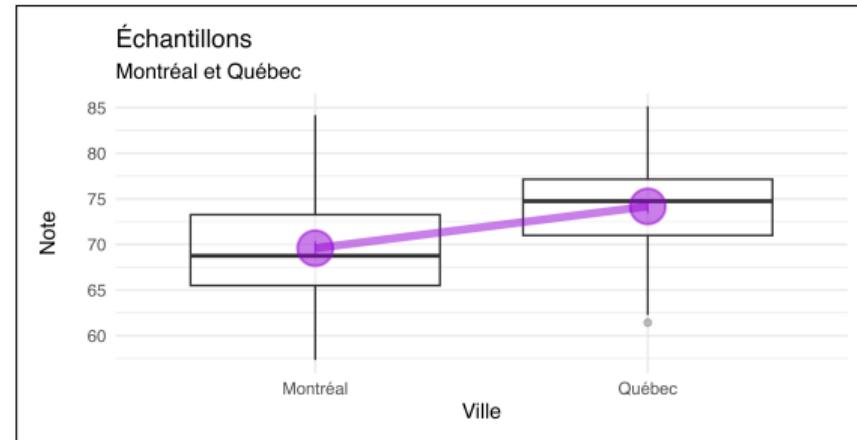
# Régression linéaire

- Nos données (dernière séance) : **variable discrète ou catégorielle**
- Essentiellement la même chose → imaginez une droite entre les deux moyennes

## Moyennes

Montréal :  $\bar{x} = 69.6$

Québec :  $\bar{x} = 74.2$



# Régression linéaire

## Synthèse de la première partie

- **Objectif** : trouver la meilleure droite pour nos données
- ☞ **L'idée principale** : la variable  $x$  a-t-elle un effet sur la variable  $y$ ?

**L'hypothèse nulle** ( $H_0$ ) : **non**, c'est-à-dire que la droite entre les villes est **horizontale**

**L'hypothèse alternative** ( $H_1$ ) : la droite n'est pas horizontale (positive ou négative)

- Comme le test  $t$  et l'ANOVA, la régression permet de **rejeter** ou de ne pas rejeter l' $H_0$



# Régression linéaire

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i_1} + \hat{e}$$

- $\hat{y}_i$  = la réponse  $i$  (la note dans notre exemple); ce qu'on veut prédire/analyser
  - $\hat{\beta}_0$  = notre constante, l'*intercept*. Donc, la valeur de  $\hat{y}_i$  quand  $\beta_1 = 0$
  - $\hat{\beta}_1$  = notre effect pour la variable **ville**, le *slope*
  - $x_{i_1}$  = l'observation  $i$  pour la variable **ville**
  - $\hat{e}$  = l'erreur (tout ce qui le modèle ne peut pas prédire)
- 
- On utilise R pour trouver  $\beta_0$  et  $\beta_1$
- ☞ L'idée sera plus claire dans quelques instants



# Régression linéaire en R

## Exécuter la régression

- Pour exécuter une régression linéaire, on utilise la fonction `lm(y ~ x)` :

```
1 # Chargez les extensions :  
2 library(tidyverse)  
3  
4 # Importer les données :  
5 villes = read_csv("Donnees/villes.csv")  
6  
7 # Visualiser les données :  
8 # ...  
9  
10 # Régression :  
11 mod1 = lm(note ~ ville, data = villes)  
12  
13 summary(mod1)
```



# Régression linéaire en R

## Interpréter le résultat

- Pour interpréter les résultats, on se concentre sur trois éléments :

```
1 Coefficients:
2   Estimate Std. Error t value Pr(>|t|)
3 (Intercept) 69.5838     0.7905  88.025 < 2e-16 ***
4 villeQuébec  4.5890     1.1179   4.105 8.38e-05 ***
5 ---
6 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
7
8 Residual standard error: 5.59 on 98 degrees of freedom
9 Multiple R-squared:  0.1467, Adjusted R-squared:  0.138
10 F-statistic: 16.85 on 1 and 98 DF,  p-value: 8.38e-05
```



# Régression linéaire en R

## Interpréter le résultat

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|) 
(Intercept) 69.5838   0.7905  88.025 < 2e-16 *** 
villeQuébec  4.5890   1.1179   4.105 8.38e-05 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.59 on 98 degrees of freedom
Multiple R-squared:  0.1467,    Adjusted R-squared:  0.138 
F-statistic: 16.85 on 1 and 98 DF,  p-value: 8.38e-05
```

1. Les coefficients ( $\hat{\beta}$ ) : l'intercept (Montréal) et le slope (Québec)
2. Les valeurs  $p$  (ici, les deux sont minuscules)
3. Le  $R^2$  : le coefficient de détermination indique la proportion des notes expliquée par `ville`



# Régression linéaire en R

## Interpréter le résultat

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|) 
(Intercept) 69.5838   0.7905  88.025 < 2e-16 *** 
villeQuébec  4.5890   1.1179   4.105 8.38e-05 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.59 on 98 degrees of freedom
Multiple R-squared:  0.1467,    Adjusted R-squared:  0.138 
F-statistic: 16.85 on 1 and 98 DF,  p-value: 8.38e-05
```

1. L'intercept nous donne la note moyenne de Montréal;  
Le slope nous donne la différence entre Montréal et Québec
2. Les valeurs  $p$  nous disent que les deux effets sont statistiquement significatifs
3. 13,8 % de la variation des notes est expliquée par la ville des apprenants



# Régression linéaire en R

## Générer les intervalles de confiance

```
1 confint(mod1)
2
3 2.5 % 97.5 %
4 (Intercept) 68.01507 71.15253
5 villeQuébec 2.37048 6.80752
```

Intervalle de confiance (IC) au niveau de 95% :

- ☞ la probabilité que la moyenne de la population ( $\mu$ ) se trouve dans l'intervalle
  - Il est important de communiquer nos intervalles ainsi que nos effets :  
les IC rendent nos résultats plus précis et plus transparent
- Règle** : si l'intervalle inclut zéro, la valeur  $p$  sera supérieure à 0,05



# Régression linéaire

## Imprimer un tableau facilement dans RStudio

```
1 # install.packages("sjPlot") # pour installer l'extension
2 library(sjPlot)
3
4 mod1 |> tab_model(title = "Tableau statistique
5 de la régression", dv.labels = "",
6 string.pred = "Variables",
7 string.ci = "IC 95%",
8 string.est = "Coefficients",
9 string.p = "Valeur p")
```

Tableau statistique de la régression			
Variables	Coefficients	IC 95%	Valeur p
(Intercept)	69.58	68.02 – 71.15	<0.001
ville [Québec]	4.59	2.37 – 6.81	<0.001
Observations	100		
R <sup>2</sup> / R <sup>2</sup> adjusted	0.147 / 0.138		



## PARTIE 2



# Régression linéaire en R

## Notre modèle

$$\hat{y}_i = \hat{\beta}_0 + \underbrace{\hat{\beta}_1 x_{i_1}}_{\text{Montréal}} + \underbrace{\hat{\beta}_1 x_{i_2}}_{\text{Québec}} + \hat{e}$$

$$\hat{y}_i = 69.58 + 4.59 \cdot 1 + \hat{e} \tag{1}$$

$$\hat{y}_i = 69.58 + 4.59 \cdot 0 + \hat{e} \tag{2}$$

1. Pour prédire la note moyenne des apprenants à Québec
2. Pour prédire la note moyenne des apprenants à Montréal

Obs. L'accent circonflexe veut dire « estimé·e »



# Régression linéaire

## Conclusion

- `ville` a un effet statistiquement significatif sur la variable `note` ( $\hat{\beta} = 4.59$ , IC 95% = [2.4, 6.8],  $p < 0.0001$ ). En plus, la ville d'un apprenant explique vers 14 % de la variation observée pour les notes des groupes. Autrement dit, la différence entre les deux groupes est réelle : on rejette l' $H_0$
- L'intercept ( $\beta_0$ ) est notre niveau de référence (Montréal)<sup>2</sup>
- Donc, c'est le slope qui montre la **différence** entre les groupes

---

<sup>2</sup>R sélectionne automatiquement le premier niveau de la variable (alphabétiquement).



# Pratique

## Une ville de plus

1. Importez le fichier `villes2.csv`
2. Explorez les différences entre les groupes par rapport à la variable `ville`
3. Créez une figure pour les variables `ville` et `note`
4. Exécutez un modèle linéaire avec ces variables
5. Interprétez les résultats
6. Répétez les étapes 3–6 avec les variables `duree` (temps d'étude du français) et `note`



# Pratique

Vrai ou faux? Justifiez votre réponse

## Révision des concepts

1. L'intercept représente la moyenne globale pour la variable  $y$
2. Si une variable  $x$  est un facteur de 6 niveaux, on aura 5 *slopes* dans notre modèle
3. Si l'intercept est 0, on accepte l'hypothèse nulle
4. L'effet d'une variable est représenté par  $\hat{\beta}$
5. La fonction pour calculer les intervalles de confiance est `conf.int()`
6. Dans une régression linéaire, les effets sont interprétés par rapport à un niveau de référence :  $\hat{\beta}_1$



# Synthèse

- On utilise une régression linéaire s'il y a une **relation linéaire** entre nos variables
- L'idée est de trouver la meilleure droite pour les données (outil interactif [ici](#))
- L'intercept est notre coefficient de référence ( $\hat{\beta}_0$ ) :  
la réponse quand les autres variables prédictives = 0
- Le slope est normalement le coefficient d'intérêt ( $\hat{\beta}_x$ )

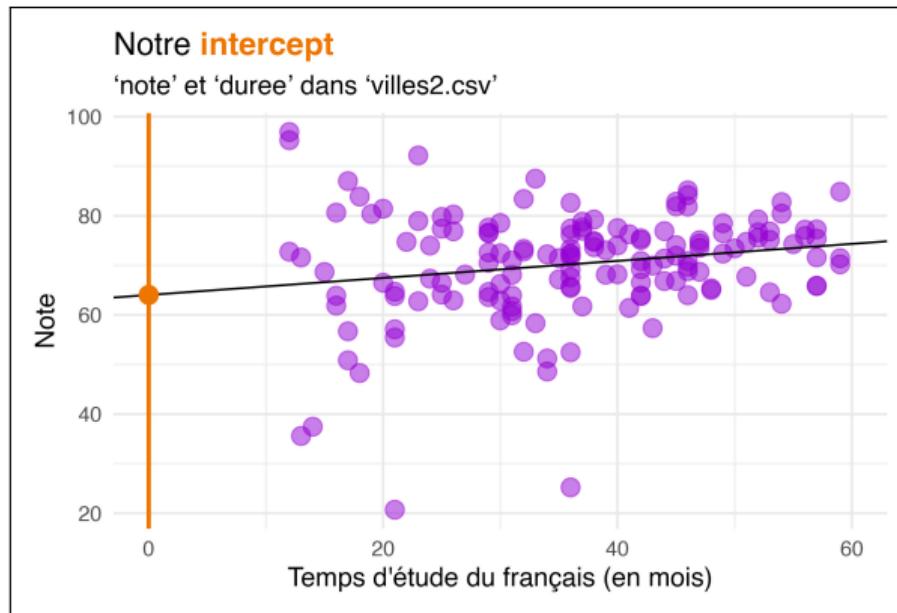
## ☞ Plus d'information :

- Garcia 2021, chap. 6 (anglais) et Barnier 2023, chap. 21 (français)



# Synthèse

- ☞ Notez que la logique sous-jacente est la même pour les variables  $x$  catégoriques ou continues



# Synthèse

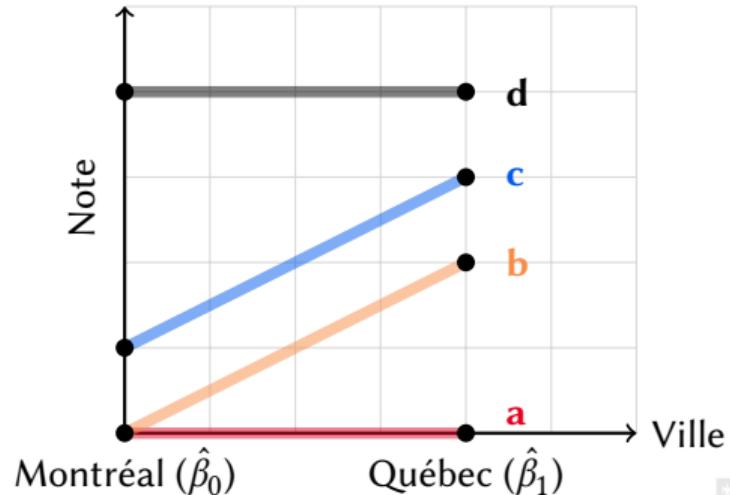
Les coefficients et l'hypothèse nulle (la variable  $x$  pourrait être continue ici)

☞  $H_0 = 0$  (on l'a déjà rejetée)

- Considérons les scénarios suivants :

Scénario	Rejetons l' $H_0$ de...	
	$\hat{\beta}_0$	$\hat{\beta}_1$
a	non	non
b	non	oui
c	oui	oui
d	oui	non

- `villes.csv` → scénario c  
(retournez à la diapo 10)



# Pratique

## Une variable de plus

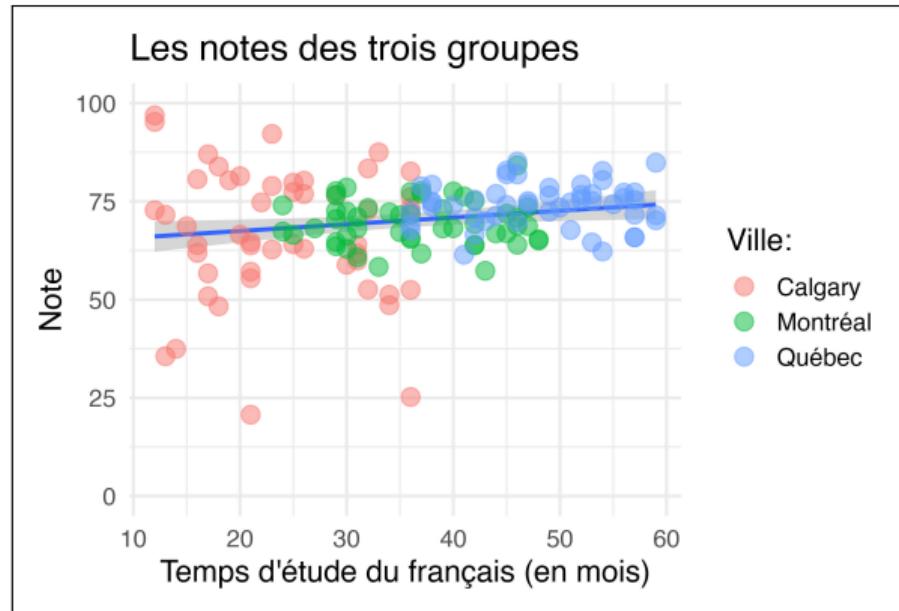
On peut avoir plusieurs variables prédictives :  $y \sim x + w + z \dots$

1. Créez une figure pour , `ville` et `duree`
2. Exécutez un modèle linéaire avec ses variables
3. Interprétez les résultats : nos conclusions changent-elles?
4. Maintenant, exécutez une régression juste avec `duree`. Qu'est-ce que vous remarquez dans les résultats?
5. **Extra** : reproduisez la figure dans la prochaine diapo (à la maison)



# Pratique

Figure à reproduire



# Notre modèle avec trois villes

Pour comprendre sa structure

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \overbrace{x_{i_1}}^{\text{Calgary}} + \hat{\beta}_2 \overbrace{x_{i_2}}^{\text{Montréal}} + \overbrace{\hat{e}}^{\text{Québec}}$$

1. Pour prédire la note moyenne des apprenants à **Calgary**
2. Pour prédire la note moyenne des apprenants à **Montréal**
3. Pour prédire la note moyenne des apprenants à **Québec**

$$\begin{aligned}\rightarrow \hat{\beta}_0 + \cancel{\hat{\beta}_1 \cdot 0} + \cancel{\hat{\beta}_2 \cdot 0} &= \hat{\beta}_0 \\ \rightarrow \hat{\beta}_0 + \hat{\beta}_1 + \cancel{\hat{\beta}_2 \cdot 0} \\ \rightarrow \hat{\beta}_0 + \cancel{\hat{\beta}_1 \cdot 0} + \hat{\beta}_2\end{aligned}$$

- Considérez le modèle `note ~ duree` et le modèle `note ~ ville`
- ☞ Réfléchissez à la différence entre une variable catégorielle et une variable continue



# Pratique à la maison

Notre modèle avec trois villes + temps d'étude

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \overbrace{x_{i_1}}^{\text{Calgary}} + \hat{\beta}_2 \overbrace{x_{i_2}}^{\text{Montréal}} + \hat{\beta}_3 \overbrace{x_{i_3}}^{\text{Québec}} + \text{temps d'étude} + \hat{e}$$

Un modèle avec les deux variables prédictives : note ~ ville + duree

## L'intercept?

- Dans le modèle ci-dessus, quelle est l'interprétation de l'intercept?



# Comparaisons multiples?

- On a vu que tous les effets doivent être comparés à l'intercept
- Mais parfois il est utile/nécessaire de comparer *tous* les niveaux d'un facteur
- Par exemple, y a-t-il une différence entre **Montréal** et **Québec**...?

## Pratique

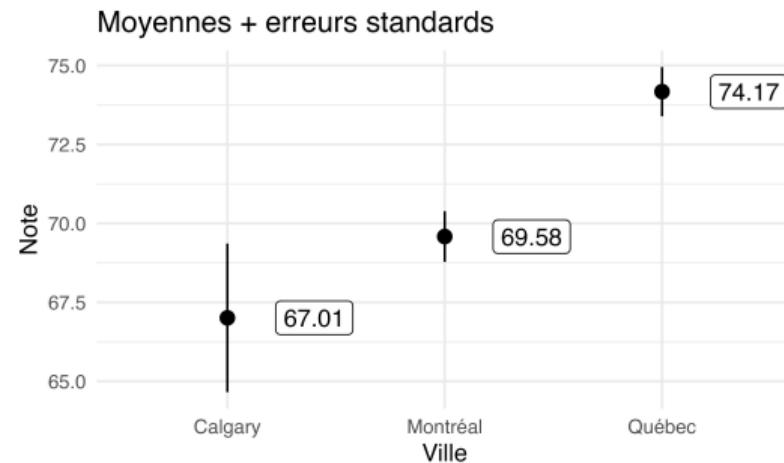
- Comment effectuer des comparaisons multiples à partir d'un régression linéaire?
- ☞ Discutez le chapitre 7 avec vos collègues et répondez aux questions du chapitre



# Extra

## L'erreur standard

- On utilise souvent l'erreur standard des moyennes dans les figures :



- Cet élément nous donne une idée de la variabilité dans les données



# Extra

## L'erreur standard

L'erreur standard est **l'écart type de la distribution d'échantillonnage**

1. Considérez une population de 20 000 personnes
2. Prenez un échantillon de 50 personnes et calculez la moyenne
3. Répétez l'étape 2 plusieurs fois (p. ex., 100), ce qui vous donne 100 moyennes

☞ L'écart type de ces moyennes est l'erreur standard de la moyenne

On peut estimer l'erreur standard manuellement :  $ES = \frac{s}{\sqrt{n}}$

code [ici](#)

Finalement, on peut calculer les intervalles de confiance au niveau de 95% :  $\bar{x} \pm 1,96 \cdot ES$



## $R^2$ vs $r$

- On a parlé du **coefficient de détermination**, le  $R^2$ . Ce coefficient  $[0, +1]$  nous donne la proportion de la variabilité expliquée dans la variable  $y$  par le modèle.
  - La **corrélation** de deux variables, le  $r$ , nous donne l'intensité de la relation entre deux variables  $[-1, +1]$ . **On ne va pas explorer ce coefficient dans notre cours.**
  - On peut arriver au  $R^2$  à partir du  $r$  ( $r^2$ ) et vice versa, mais quand on calcule la corrélation à partir du  $R^2$  on n'a pas la direction de la corrélation ( $r = \sqrt{R^2}$ ).
- ☞ Donc, une corrélation de  $-0.5$  ou de  $0.5$  nous donne un  $R^2$  de  $0.25$  (la direction de la corrélation est perdue dans le  $R^2$ )



# Références I

Julien BARNIER : *Introduction à R et au tidyverse*. 2023. Available at <https://juba.github.io/tidyverse/>.

Guilherme D GARCIA : *Data visualization and analysis in second language research*. Routledge, New York NY, 2021.

