

---

# LNG-1100 : Méthodes expérimentales et analyse de données

Analyse de données : concepts basiques

Guilherme D. Garcia

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

4



UNIVERSITÉ  
LAVAL

# Plan de la séance

Dans RStudio aujourd'hui

1. Réviser la visualisation de données avec l'extension `ggplot2`
2. Intro à l'analyse de données
  - échantillonnage et population
  - simulation des données
  - valeur  $p$
  - test  $t$  (exemple)
3. Pratique

# Visualisation de données

## Pratique (ensemble)

Observez les différents types de graphiques [ici](#)

1. Quelles fonctions de l'extension `ggplot2` sont utilisées ici?
2. Comment nos conclusions sont-elles affectées par nos choix de visualisation?

# Visualisation de données

code

```
1      # Tableau sampleData.csv (format long) :  
2      participant group   test   note  
3      <chr>         <chr>  <chr> <dbl>  
4      1 subject_1   control testA   4.4  
5      2 subject_1   control testB   6.9  
6      3 subject_1   control testC   6.3  
7      4 subject_2   control testA   6.5  
8      ...
```

## Pratique

1. Comment créer un graphique de boîte à moustache à partir du tableau ci-dessus?
2. Comment créer un graphique de moyennes + barres d'erreur?

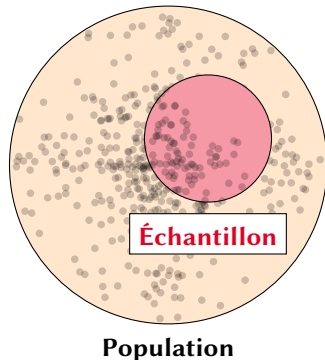
# L'ANALYSE DE DONNÉES

# Échantillon vs population

- « Analysez les notes des apprenants de français à Québec »
- Supposez qu'il y en a 20 000 (la **population** complète)
- Chaque apprenant a complété un test de français (0–100)

☞ C'est **trop!** Donc, on en prélève un petit **échantillon**  
*On déduit la population à partir de l'échantillon*

**Cette technique est-elle précise...?**



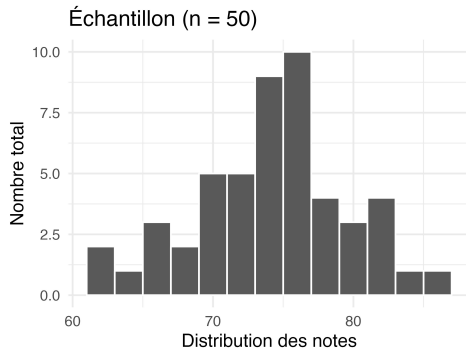
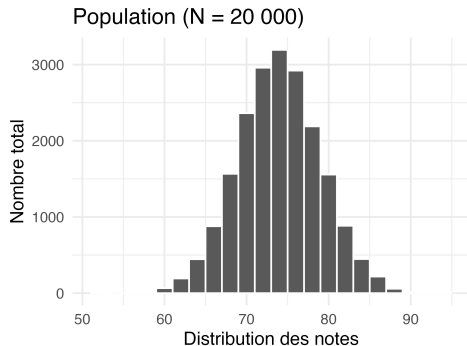
# Simulation de données

code

```
1      # Simuler 20 000 notes : la population
2      set.seed(1)
3      population = rnorm(n = 20000, mean = 74, sd = 5)
4
5      # Vérifier la moyenne des données simulées :
6      mean(population) # 73.97318 (suffisamment proche!)
7
8      # Prélever un échantillon de 50 participants de la population :
9      set.seed(2)
10     echant = sample(x = population,
11                    size = 50,
12                    replace = FALSE) # On ne répète pas les participants
13
14     mean(echant) # 74.17266 : très précis!
15
16
17     # set.seed(...) permet de reproduire les résultats
```

# Simulation de données

- Voici la distribution des notes : la population présente une distribution normale<sup>1</sup>



<sup>1</sup>Ou une distribution [gaussienne](#)



# Simulation de données

Important!

- On n'a **jamais** accès direct à la population<sup>2</sup> : on n'examine qu'un **échantillon**
- C'est une des raisons pour lesquelles on utilise la statistique inférentielle :  
on « estime le tout à partir de la partie »

---

<sup>2</sup>Sauf si on parle d'une population minuscule!

# Un autre groupe d'apprenants

- Maintenant, on veut examiner les apprenants de français à Montréal
- Cette fois-là, on a simplement un échantillon de 50 participants
- ☞ Donc, on va comparer 50 participants de Québec et 50 participants de Montréal

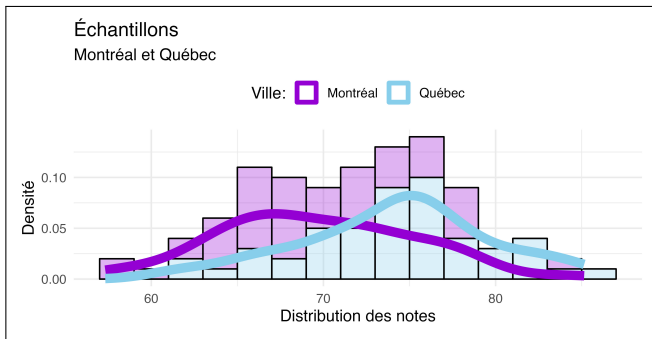
# Un autre groupe d'apprenants

## Pratique en groupes (10 min)

1. Importez le fichier `villes.csv` (monPortail)
2. Calculez la note moyenne pour chaque groupe de participants
3. Ordonnez les notes en ordre décroissant
4. Exportez le tableau en tant que `villesOrdonnees.csv`
5. Créez un graphique pour comparer les deux groupes

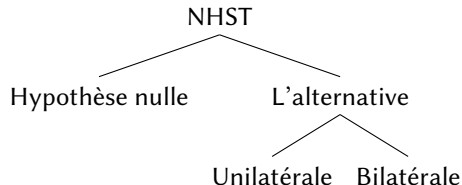
# Comparaison

- Moyennes ( $\bar{x}$ ) des apprenants : Québec  $\bar{x} = 74.2$  Montréal  $\bar{x} = 69.6$
  - La question est si on peut conclure que les deux groupes sont **réellement différents**
- 👉 **Autrement dit** : les deux échantillons viennent-ils des deux populations **différentes**?



# Comparaison

- **Hypothèse nulle** ( $H_0$ ) :  
il n'y a aucune différence réelle entre les groupes (même population sous-jacente)



- La façon la plus simple d'analyser nos données : [test  \$t\$](#) , un test **paramétrique**<sup>3</sup>

---

<sup>3</sup>Pour les données qui suivent la **loi normale** : voici un [short](#) et une [video](#).

# Test $t$

code

```
1      > mtl_qb
2      # A tibble: 100 × 2
3      note ville
4      <dbl> <chr>
5      1  65.9 Québec
6      2  64.6 Québec
7      3  74.8 Québec
8      4  69.7 Québec
9
10     > t.test(note ~ ville, data = mtl_qb)
11
12     Welch Two Sample t-test
13
14     data:  note by ville
15     t = -4.1046, df = 97.919, p-value = 8.392e-05
16     alternative hypothesis: true difference in means between group Montréal and group Québec is not equal to
17     95 percent confidence interval:
18     -6.807334 -2.370219
19     sample estimates:
20     mean in group Montréal    mean in group Québec
21     69.58388                74.17266
```

# Test $t$

## Interprétation

- 10. Syntaxe pour la fonction :  $\text{note} \sim \text{ville}$   
c.à.d « l'analyse de la note **en fonction de** la ville »
- 12. Le type de test : *two sample* = bilatérale  
« l'un ou l'autre des échantillons pourrait avoir une moyenne plus élevée »
- 15.  $t = -4.1$  notre statistique (valeur  $t$ ) et notre valeur  $p = 8.4^{-5} = 0.000084$
- 18. L'intervalle de confiance au niveau de 95%
- 20. Les moyennes des deux échantillons

# Test $t$

## Interprétation

### Résumé :

- Notre test  $t$  indique que les deux groupes sont statistiquement différents ( $p < 0.0001$ )
- La probabilité d'observer la différence en question si l'hypothèse nulle est correcte est **minuscule** (notre seuil de décision est typiquement 5%, soit 0.05)

👉 Donc, les apprenants de Québec ont la moyenne la plus élevée



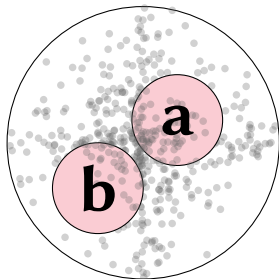
# Test $t$

## Pratique

1. Dans le fichier `villes.csv`, sélectionnez les notes supérieures à 60
2. Créez un graphique de boîte à moustaches
3. Exécuter un test  $t$  et interprétez les résultats
4. Lisez la documentation de la fonction `?t.test` et explorez l'argument `alternative`

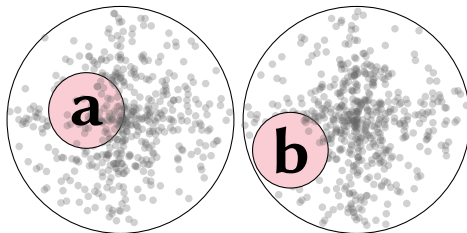
# Résumé

👉 **Hypothèse nulle** : **a** et **b** ne sont pas différents; ils viennent de la **même population** ( $p \geq 0,05$ ). Autrement dit,  $\mu_a = \mu_b$ .



Population

👉 **Hypothèse alternative** : **a** et **b** sont différents; ils viennent des populations **différentes** ( $p < 0,05$ ). Autrement dit,  $\mu_A \neq \mu_B$ .




Population A

Population B

# Test $t \rightarrow$ ANOVA

- Il y a plusieurs limitations dans les tests  $t$ . Par exemple :
  - Seulement deux groupes peuvent être comparés
  - Seulement un variable peut être incluse dans l'analyse (ville ici)
  - ...

# Semaine prochaine

- **ANOVA** : Lisez attentivement [ch. 5 du livre du cours](#) 
- Faites les exercices et consultez les solutions du chapitre avant la séance

(Garcia 2024, ch. 5)

# Références I

Guilherme D GARCIA : Méthodes expérimentales et analyse de données. <https://lng1100.quarto.pub/>, 2024. Livre numérique du cours LNG1100 de l'Université Laval.