LNG-1100 : Méthodes expérimentales et analyse de données

Analyse de données : test $t \rightarrow$ ANOVA

Guilherme D. Garcia

fr.gdgarcia.ca

5

Pratique

1. Dans quelles conditions pouvons-nous utiliser le test t?
2. Quelle est la fonction et la syntaxe pour exécuter le test?
3. Quelle est l'hypothèse nulle dans un test t?
4. Quelle est l'interprétation de la valeur p?
5. Qu'est-ce qu'on voit dans le résultat d'un test t?
6. Quelles sont les limitations du test t?

Révision

Test *t*

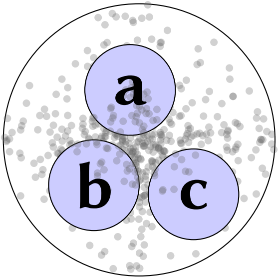
Pratique

1. Pour vérifier si deux groupes sont statistiquement différents par rapport à une variable **continue**, ou pour vérifier si un groupe est statistiquement différent d'une valeur spécifique (normalement, zéro).
2. `t.test(y ~ x, data = ...)`
3. Que les moyennes des groupes sont identiques : $\mu_A = \mu_B$ ou que la moyenne d'un groupe est identique à zéro (ou à la valeur définie par l'utilisateur)
4. C'est la probabilité de voir les données (et la différence) en question si l'hypothèse nulle est vraie
5. La valeur *p*, l'intervalle de confiance à 95 %, les moyennes pertinentes, les dégrées de liberté, la valeur *t*
6. 1 ou 2 groupes; juste une variable analysable (groupe); très simpliste

Quand il y a plus de 2 groupes

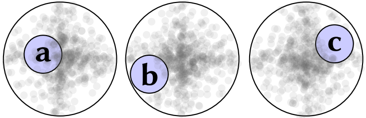
Et plus d'une variable...?

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



Population

☞ **Hypothèse alternative** : **a**, **b** et **c** sont différents; ils viennent des populations **différentes** ($p < 0,05$). Autrement dit, $\mu_a \neq \mu_b \neq \mu_c$.



Pop A **Pop B** **Pop C**

Les limitations du test *t*

- On a souvent plus de deux groupes dans notre analyse
- En plus, on veut analyser plusieurs variables en même temps

☞ Un test *t* est simplement **trop limité**

Aujourd'hui : **ANOVA** (*ANalysis Of VAriance*)

- Une méthode qui nous permet d'analyser plusieurs groupes/variables en même temps
- Ici, l'ANOVA sera examinée de façon temporaire : on cible les régressions complètes

☞ Mais il est important de bien connaître l'ANOVA :
la littérature contient beaucoup d'articles que l'utilisent

$$\text{test } t \longrightarrow \text{ANOVA} \longrightarrow \boxed{\textbf{régressions}}$$

Pratique

Révision du chapitre 5

Questions de base

1. Si on examine 5 villes dans notre fichier, combien de tests t faudra-t-il exécuter pour comparer toutes les villes?
2. Quel est le problème de cet approche?
3. Quels sont les deux types d'erreurs pertinents à l'analyse de données?

Pratique

Révision du chapitre 5

Questions de base

1. 10 (AB, AC, AD, AE, BC, BD, BE, CD, CE, DE) $\frac{k(k-1)}{2}$
2. Le taux d'erreur de **type 1** explose $1-(1-\alpha)^{10}$
3. Type 1 (faux positif) et type 2 (faux négatif)

Erreurs

Nos possibilités

	H_0 est vraie	H_0 est fausse
on rejette H_0	Type I	
on ne rejette pas H_0		Type II

Exemple classique : un test de **grossesse**

- **Positif** mais la femme n'est pas enceinte — erreur de type I
- **Négatif** mais la femme est enceinte — erreur de type II

☞ Ces concepts sont pertinent n'importe quelle méthode on utilise dans le cours

ANOVA

Concepts de base

- L'idée générale : $F = \frac{\text{variabilité \textbf{entre} les groupes}}{\text{variabilité à l'\textbf{intérieur} des groupes}}$

☞ Si $F > 1$, peut-être les groupes sont différents

Vrai ou faux?

1. Une ANOVA nous montre où sont les différences entre des groupes
2. La fonction utilisé pour exécuter une ANOVA est `anova()`
3. L'hypothèse alternative (H_1) d'une ANOVA est que tous les groupes sont différents

ANOVA

Correction

Vrai ou faux?

1. Une ANOVA nous montre où sont les différences entre des groupes **FAUX**
2. La fonction utilisé pour exécuter une ANOVA est `anova()` **FAUX**
3. L'H_1 d'une ANOVA est que tous les groupes sont différents **FAUX**

Qu'est-ce la variance...?

La variabilité à l'intérieur des groupes

- Si l'ANOVA cible la variance, il faut bien comprendre la définition de **variance**
- Voici les premières lignes de `villes2.csv` (version simplifiée)

☞ Analysons notre tableau : pour Calgary, la note moyenne est 67.

	note	ville
1	52.47	Calgary
2	68.67	Calgary
3	48.29	Calgary
4	96.91	Calgary
5	71.59	Calgary
6	48.59	Calgary
...	...	...

Qu'est-ce la variance...?

La variabilité à l'intérieur des groupes

- On calcule la **différence** (l'écart) entre chaque note et la moyenne du groupe
- Après, on calcule le **carré** de l'écart (ce qui nous donnera juste des valeurs positives)

	note	ville	moyenne	note-moyenne	(note-moyenne)^2
1	52.47	Calgary	67.00	-14.53	211.09
2	68.67	Calgary	67.00	1.67	2.80
3	48.29	Calgary	67.00	-18.71	350.16
4	96.91	Calgary	67.00	29.91	894.35
5	71.59	Calgary	67.00	4.59	21.07
6	48.59	Calgary	67.00	-18.41	338.90
...	...	...	...	...	...
					$\boxed{\frac{somme}{N-k}}$

- La variance à l'intérieur des groupes sera la $\boxed{\textbf{somme totale divisée par } N-k}$

  N = nombre total d'observations; k = nombre de groupes (villes ici)

ANOVA

Concepts de base

- L'idée générale : $F = \frac{\text{variabilité } \textbf{entre} \text{ les groupes}}{\text{variabilité à l'}\textbf{intérieur} \text{ des groupes}}$

☞ Maintenant, calculons la variabilité **entre** les groupes

Qu'est-ce la variance...?

La variabilité entre les groupes

- Calculez les **moyennes** par groupe ainsi que la **différence** entre leurs moyennes et la moyenne générale ($\bar{x} = 70.25$)
- Après, multipliez les carrés des écarts (CE) par le nombre d'observations (n) $\rightarrow$ n_CE

	ville	moyenne	n	diff	CE	n_CE
1	Calgary	67.01	50	-3.25	10.54	526.90
2	Montréal	69.58	50	-0.67	0.45	22.54
3	Québec	74.17	50	3.92	15.35	767.38
						$\boxed{\frac{somme}{k-1}}$

- La $\boxed{\text{somme totale}}$ est donc divisée par $k-1$, et cela sera notre variété entre les groupes

Pratique

Chapitre 5

- Examinons les code et les exercices dans le chapitre

ANOVA + R

- Heureusement, il y a une fonction qui automatise le calcul pour nous : `aov()`

Pratique

Complétez le script `seance-5.R` :

1. Calculez les moyennes et les écarts-types des cinq groupes.
2. Visualisez les données et exécutez une ANOVA.
3. Pouvons-nous rejeter l'hypothèse nulle? Générez des comparaisons multiples.
4. Avons-nous des erreurs dans les résultats?
5. Communiquer les résultats en utilisant le modèle présenté dans le chapitre.

ANNEXE : LA DISTRIBUTION F

La distribution F

Pour mieux comprendre la logique de l'ANOVA

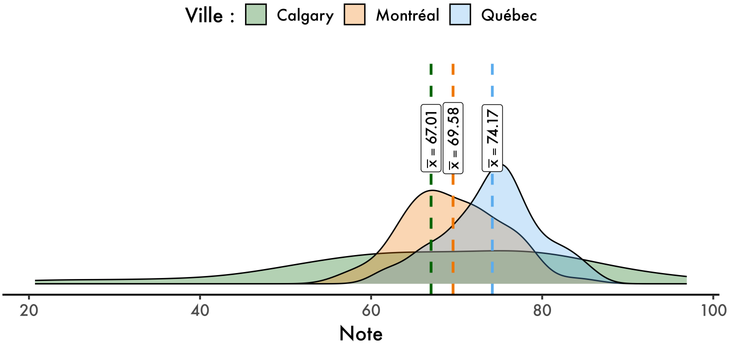
1. On calcule les variabilités et la valeur F
2. Ensuite, avec les **dégrées de liberté** des données (**2** et **147** pour `villes2.csv`),[1] on consulte un tableau de valeurs critiques. Pour $\alpha = 0.05$ et une hypothèse bilatérale cette valeur sera de ≈ 3.06. Donc, si notre valeur F est supérieure à cette valeur, on sera dans la **région critique**, ce qui nous permettra de rejeter l'hypothèse nulle.

☞ Examinez le tableau en question : quelle est la relation entre les dégrées de liberté et les valeurs critiques de F?

[1] Nombre de groupes (3) - 1. Nombre d'observations (150) - nombre de groupes (3).

La distribution F

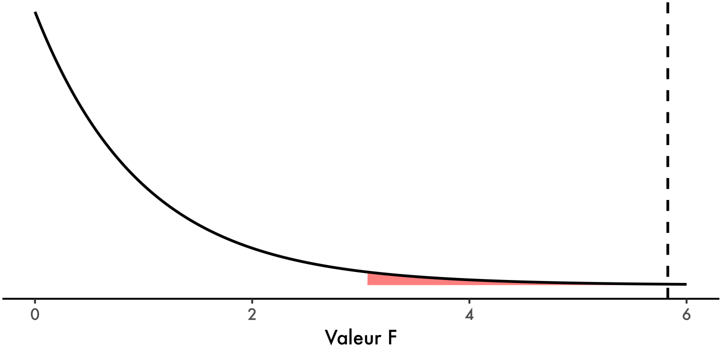
Visualisons nos données



- Après avoir calculé F (ou après avoir exécuté `aov(...)`), on arrive à $F = 5.82$

La distribution F

La région critique (rouge) = 5 % de la distribution



☞ 5.82 (ligne pointillée) est **beaucoup** plus élevé que 3.06, la valeur critique pour $F(2, 147)$

- On est donc dans la région critique → on **rejette** l'hypothèse nulle

Commentaires finaux

Test *t* vs. ANOVA

- L'ANOVA et le test *t* suppose que la variable de réponse est **normale**
- En plus, les deux méthodes suppose que la variance est la même à travers les groupes
- Le test *t* est limité à 1 ou 2 groupes; l'ANOVA est libre
- L'ANOVA peut avoir plusieurs variables : `aov(y ~ x + w + z)`
- Ces méthodes nous donnent une valeur *p*, ce qui nous permet de rejeter ou de ne pas rejeter l'hypothèse nulle. La statistique nous permet de conclure si un effet ou si une différence est **significative** ou **crédible**. Notre analyse ne doit pas pourtant se concentrer simplement sur les valeurs *p*!

☞ Les deux méthodes servent de point de départ pour le cours