Seance 10: Comparaison de moyenne

Visseho Adjiwanou, PhD.

30 March 2022

Rappel

1. Comparaison de la moyenne (Chapitre 8 de Fox)
2. Test de comparaison de la moyenne: test t de student

			Type de la variable dépendante		
			Catégorielle		Continue
			Nominale	Ordinale	Intervalle/ratio
Type de la variable indépendante	Catégorielle	Nominale	Tableau croisé Test du chi-carré V de Cramer Diagramme de barre empillée	Tableau croisé	Analyse de la variance Test t État carré Diagramme boxplot
		Ordinale	Tableau croisé	Tableau croisé	idem
	Continue	Intervalle/ratio	Transformer la VI en catégorielle		Corrélation / régression Test t R carré Diagramme de nuage de points

Figure 1

Comparaison de moyenne

- Exemple
- Différence des heures d'écoute de la télé entre Montréalais et Torontois
- Différence de revenus entre les hommes et les femmes
- Autres exemples?

Comparaison de moyenne

- Comme on le voit, on est en présence de deux variables:
- Une variable dépendante (intervalle/ratio)
  - Heure d'écoute de la télé
  - Revenus
- Une variable indépendante dichotomique
  - Résidence (Montréal, Toronto)
  - Sexe (Homme, Femme)

Comparaison de moyenne

1. Représentation

- Diagramme en boîte (à moustache) – boxplot
- Mais, attention, il ne nous donne pas la moyenne
- Il faut ajouter la moyenne de chaque groupe sur le graphique

2. Calcul

- Très simple: calculé la moyenne de la variable dépendante dans chaque catégorie de la variable indépendante

Comparaison de moyenne

- Comment savons-nous que la différence observée dans l'échantillon est valide au sein de la population?
- Il faut faire un **test statistique** pour conclure.
- Test est basé sur la distribution d'échantillonnage de la différence de la moyenne

Comparaison de moyenne

- Prendre tous les échantillons possible de taille N de la population
- Calculer la moyenne pour les deux catégories
- Prendre la différence de ces moyennes: vous vous retrouvez avec des millions de différences

Test t

1. S'il y a **une grande différence** au sein de la **population**

- On va s'attendre à ce que dans la très grande majorité des échantillons possibles, la différence est aussi grande
- Quelque chose comme dans 19 fois sur 20, la différence va être significative, ou que 1 fois sur 20, elle ne va pas l'être

2. En revanche, s'il **n'y a pas une grande différence** au sein de la **population**

- On va s'attendre à ce que dans la très grande majorité des échantillons possibles, la différence est aussi faible, autour de 0

Test t

1. Plus la différence entre les moyennes est faible dans la population, plus grande est la proportion des échantillons ayant une faible différence entre les moyennes
2. Plus la différence entre les moyennes est forte dans la population, plus grande sera la proportion des échantillons ayant une forte différence entre les moyennes.

Distribution d'échantillonnage

- Prendre tous les échantillons possible de taille N de la population
- Calculer la moyenne pour les deux catégories

Distribution d'échantillonnage

- Prendre tous les échantillons possible de taille N de la population
- Calculer la moyenne pour les deux catégories
- Prendre la différence de ces moyennes: vous vous retrouvez avec des millions de différences

Distribution d'échantillonnage

- Prendre tous les échantillons possible de taille N de la population

- Calculer la moyenne pour les deux catégories
- Prendre la différence de ces moyennes: vous vous retrouvez avec des millions de différences
- La représentation graphique de cette différence est la **distribution d'échantillonnage** de la différence

Distribution d'échantillonnage

- Prendre tous les échantillons possible de taille N de la population

- Calculer la moyenne pour les deux catégories
- Prendre la différence de ces moyennes: vous vous retrouvez avec des millions de différences
- La représentation graphique de cette différence est la **distribution d'échantillonnage** de la différence
- Les statisticiens ont montré que cette distribution tend vers une **loi normale** au fur et à mesure que N devient grand

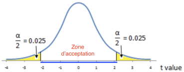
Distribution d'échantillonnage

- Prendre tous les échantillons possible de taille N de la population
- Calculer la moyenne pour les deux catégories
- Prendre la différence de ces moyennes: vous vous retrouvez avec des millions de différences
- La représentation graphique de cette différence est la **distribution d'échantillonnage** de la différence
- Les statisticiens ont montré que cette distribution tend vers une **loi normale** au fur et à mesure que N devient grand
- Plus concrètement, cette distribution suit une **loi de Student**

Distribution de Student et lecture

Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	**1-Tail Confidence Level**
	80%	90%	95%	98%	99%	99.9%	**2-Tail Confidence Level**
	0.100	0.050	0.025	0.010	0.005	0.0005	**1-Tail Alpha**
df	0.20	0.10	0.05	0.02	0.01	0.001	**2-Tail Alpha**
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896	
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739	
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594	
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460	

Figure 2

Test t

Cette statistique s'écrit :

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

Test t

$\bar{X}_1$ = la moyenne de la variable dépendante pour la catégorie 1 de la variable indépendante de l'**échantillon**

$\bar{X}_2$ = la moyenne de la variable dépendante pour la catégorie 2 de la variable indépendante de l'**échantillon**

μ_1 = la moyenne de la variable dépendante pour la catégorie 1 de la variable indépendante de la **population**

μ_2 = la moyenne de la variable dépendante pour la catégorie 2 de la variable indépendante de la **population**

$s_{\bar{X}_1 - \bar{X}_2}$ = erreur-type de la différence entre les moyennes

Test t

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{(\frac{N_1 * s_1^2 + N_2 * s_2^2}{N_1 + N_2 - 2})(\frac{N_1 + N_2}{N_1 * N_2})}$$

$s_1^2 =$ variance de la variable dépendante pour la catégorie 1 de la variable indépendante

$s_2^2 =$ variance de la variable dépendante pour la catégorie 2 de la variable indépendante

$N_1 =$ Le nombre de cas dans la catégorie 1 de la variable indépendante

$N_2 =$ Le nombre de cas dans la catégorie 2 de la variable indépendante

Exemple de calcul

Hommes (Catégorie 1)	Femmes (Catégorie 2)
$\bar{X}_1 = 2.75$	$\bar{X}_2 = 3.01$
$s_1 = 2.030$	$s_2 = 2.225$
$N_1 = 855$	$N_1 = 1085$

Test statistique

1. Hypothèse nulle : $\mu_1 = \mu_2$ (la moyenne dans le groupe 1 de la population est la même que dans le groupe 2)

   - Ou Hypothèse nulle : $\mu_1 - \mu_2 = 0$

2. Postule qu'il n'y a pas de différence entre la moyenne pour les femmes et la moyenne pour les hommes dans la population
3. Dans ces conditions, le t devient $t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{X}_1 - \bar{X}_2}}$
4. Le degré de liberté vaut $N_1 + N_2 - 2$
5. Nous décidons de rejeter ou de ne pas rejeter H_0, en déterminant où se trouve cette valeur dans la distribution t

En résumé (1e manière)

Pour faire un test de comparaison de deux moyennes:

1. Posez votre hypothèse nulle
2. Choisissez votre niveau de significativité
3. Trouvez votre degré de liberté
4. Trouvez votre zone d'acceptabilité et de rejet (où trouver la valeur MINIMALE de t pour rejeter l'hypothèse nulle)
5. Calculez votre t selon les formules ci-dessus
6. Prendre une décision:

- Si votre $|t|$ calculé est supérieur ou égal au t(lu) $==>$ Rejeter l'hypothèse nulle
- Si votre $|t|$ calculé est inférieur au t (lu) $==>$ Vous ne pouvez pas rejeter l'hypothèse nulle

En résumé (2e manière)

OU

6. Prendre une décision:

   - Si votre t calculé se trouve entre la zone d'acceptation : Ne rejeter pas l'hypothèse nulle
   - Si votre t calculé se trouve à l'extérieur de la zone d'acceptation : rejeter l'hypothèse nulle

Exemple

- Est ce que les hommes regardent la télé plus que les femmes?
- Vous obtenez les résultats suivants à partir d'un échantillon de . . . (à vous de trouver la valeur) hommes et femmes

Hommes (Catégorie 1)	Femmes (Catégorie 2)
$\bar{X}_1 = 2.75$	$\bar{X}_2 = 3.01$
$s_1 = 2.030$	$s_2 = 2.225$
$N_1 = 855$	$N_1 = 1085$

Exemple (1e manière)

1. Posez votre hypothèse nulle
   - H_0 : La moyenne d'heures d'écoute de la télé est la même pour les hommes et les femmes

Exemple (1e manière)

1. Posez votre hypothèse nulle

- H_0 : La moyenne d'heures d'écoute de la télé est la même pour les hommes et les femmes

2. Choisissez votre niveau de significativité

- $\alpha = 0.05$

Exemple (1e manière)

1. Posez votre hypothèse nulle
   - H_0 : La moyenne d'heures d'écoute de la télé est la même pour les hommes et les femmes
2. Choisissez votre niveau de significativité
   - $\alpha = 0.05$
3. Trouvez votre degré de liberté
   - df = 855 + 1085 - 2 = 1938

Exemple (1e manière)

1. Posez votre hypothèse nulle
   - H_0 : La moyenne d'heures d'écoute de la télé est la même pour les hommes et les femmes
2. Choisissez votre niveau de significativité
   - $\alpha = 0.05$
3. Trouvez votre degré de liberté
   - df = 855 + 1085 - 2 = 1938
4. Trouvez la valeur MINIMALE de t pour rejeter l'hypothèse nulle
   - t(lu) = 1.960

Exemple (1e manière)

5. Calculez votre t selon les formules ci-dessus

Rappelons-nous la formule de t:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left(\frac{N_1 * s_1^2 + N_2 * s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 + N_2}{N_1 * N_2}\right)}$$

```
numerateur <- 2.75 - 3.01
denominateur <- sqrt((855*2.030^2 + 1085*2.225^2)/(855 + 1085
t_calcule <- numerateur/denominateur

t_calcule
```

```
## [1] -2.653868
```

Exemple (1e manière)

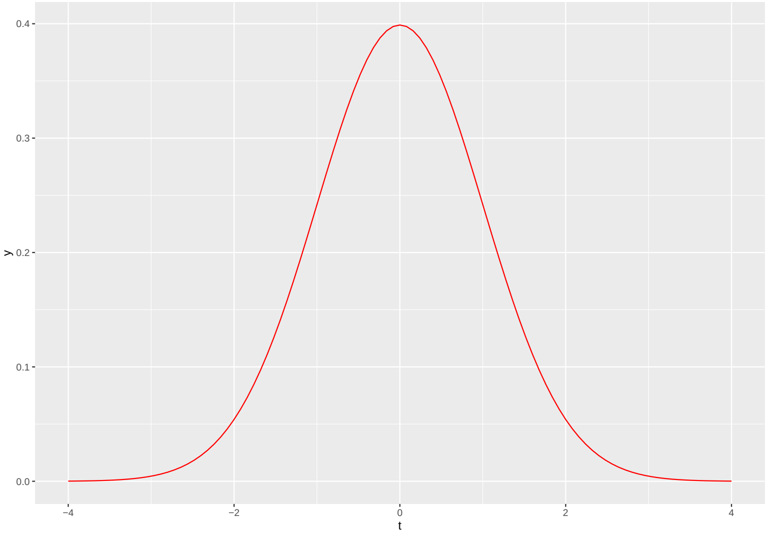
6. Prendre une décision:
   - Si votre |t| calculé est supérieur ou égal au t(lu) ==> Rejeter l'hypothèse nulle
   - Si votre |t| calculé est inférieur au t (lu) ==> Vous ne pouvez pas rejeter l'hypothèse nulle

- Notre décision
  - $|-2.653868| = 2.653868 > 1.960$,
  - décision: On rejette l'hypothèse nulle.

Exemple (2e manière)

- H_0 : La moyenne d'heures d'écoute de la télé est la même pour les hommes et les femmes
- Sous cette hypothèse, la statistique t prend la valeur : $t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{X}_1 - \bar{X}_2}}$
- Le degré de liberté vaut $N_1 + N_2 - 2 = 855 + 1085$ - $2 = 1938$
- Avec ce niveau de degré de liberté, la distribution de Student aura l'allure ci-dessous:

Exemple



Exemple

- Si dans la population, il n'y a pas relation entre le sexe et la durée d'écoute de la télé (Hypothèse nulle), on s'entend que si on tire par exemple, 20 échantillons aléatoires de taille N, la grande majorité va avoir la valeur de t autour de 0.

Exemple

- Si dans la population, il n'y a pas relation entre le sexe et la durée d'écoute de la télé (Hypothèse nulle), on s'entend que si on tire par exemple, 20 échantillons aléatoires de taille N, la grande majorité va avoir la valeur de t autour de 0.
- Si nous souhaitons nous tromper 1 fois sur 20 (niveau de significativité de 5%), alors, il faut trouver :
    - la borne inférieure et
    - la borne supérieure

Exemple

- Si dans la population, il n'y a pas relation entre le sexe et la durée d'écoute de la télé (Hypothèse nulle), on s'entend que si on tire par exemple, 20 échantillons aléatoires de taille N, la grande majorité va avoir la valeur de t autour de 0.
- Si nous souhaitons nous tromper 1 fois sur 20 (niveau de significativité de 5%), alors, il faut trouver :
  - la borne inférieure et
  - la borne supérieure
- A partir de cette courbe, on peut ainsi trouver la valeur de t pour laquelle
  - la distribution est inférieure à 0,025 (valeur inférieure)
  - la distribution est inférieure à 0,975 (valeur supérieure)

Exemple

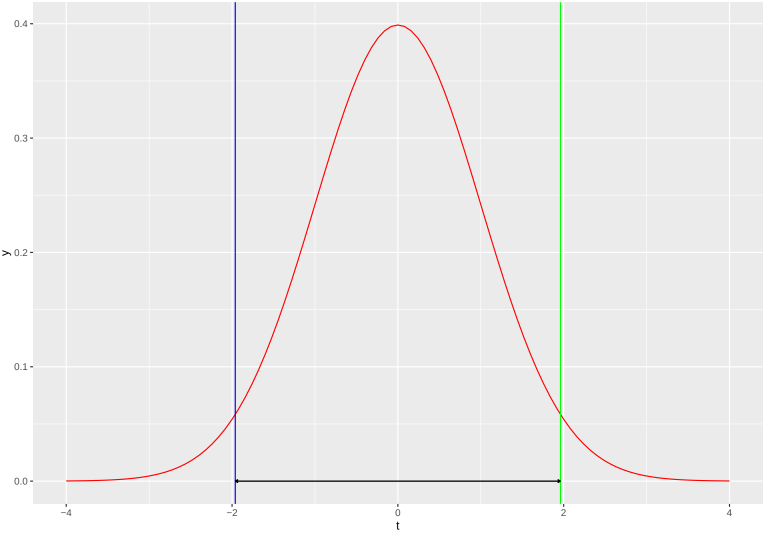
```r
bi <- qt(0.025, df=1938)
bs <- qt(0.975, df=1938)

c(bi, bs)
```

```
## [1] -1.961189  1.961189
```

Plaçons ces informations sur le graphique

Plaçons ces informations sur le graphique



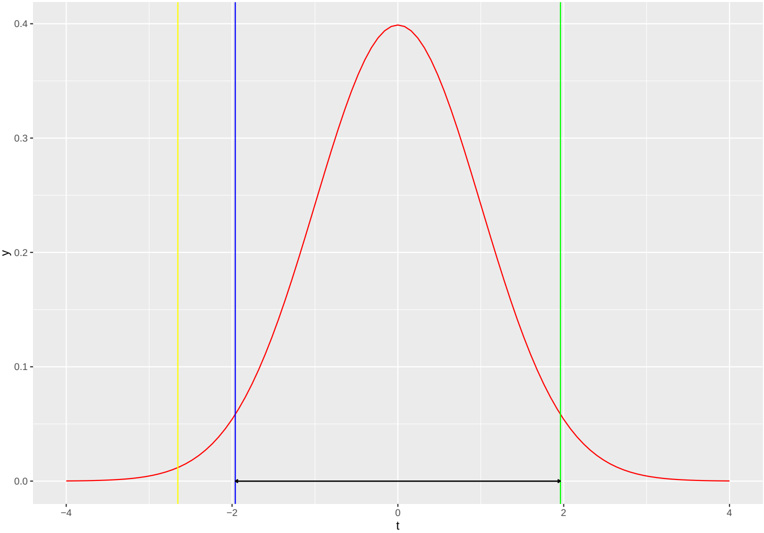
Décision

- Maintenant, si vous calculer votre t issu de votre échantillon et qu'il se trouve
  - entre -1,96 et 1,96 (entre les lignes bleue et verte), alors vous ne pouvez pas rejeter votre hypothèse nulle
  - plus petit que -1,96 ou plus grand que 1,96, alors on rejette l'hypothèse nulle.

Calculons alors le t

- On avait calculé le t qui vaut: -2.653868
- On voit ainsi que la valeur du t calculé se trouve en dehors de la zone d'acceptation:
- Conclusion: On rejette l'hypothèse nulle.
- En classe, je vais vous montrer comment lire la table de distribution du t de Student

Décision



Conclusion

- Pour la semaine prochaine
  - Faire Labo 10 sur la comparaison de moyenne
  - Lire chapitre 9
- Le Quiz va porter sur tout ce qu'on a vu jusqu'au chapitre 8