

GC – Probabilités et Statistique

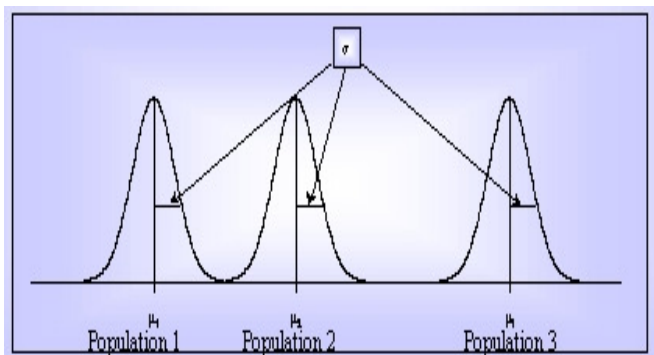
<http://moodle.epfl.ch/course/view.php?id=14271>

Cours 11a

- 1-way ANOVA (anova à une voie)
- Évaluation du modèle (*vidéo seulement – NON-EXAMINÉE*)
- Comparaisons multiples (*vidéo seulement*)

ANOVA

- Abréviation de *AN*alysis *Of* *VA*riance (analyse de variance)
- Mais c'est un test de différences des *moyennes*
- L'idée :



Principe du test

- L'analyse de variance à un facteur teste l'effet d'un *facteur* A ayant k modalités sur les moyennes d'une *variable quantitative* X
- Les hypothèses testées sont les suivantes :

$$H : \mu_1 = \mu_2 = \dots = \mu_k = \mu \text{ contre } A : \exists \mu_i \neq \mu_j$$

- Tester si le rapport des 2 estimateurs de variance est proche de 1
- Les estimations des variances associées ou *carré moyen* sont :
 - *Variance totale* : $SCE_{totale}/(n-1)$
 - *Variance due au facteur* A (CM_{trts}) : $SCE_{trts}/(k-1)$
 $\implies$ estimateur de σ^2 si H est vraie
 - *Variance résiduelle* (CM_{erreur}) : $SCE_{erreur}/(n-k)$
 $\implies$ estimateur de σ^2 quelque soit le modèle

Les modèles

- $\epsilon_{ij} \sim \text{iid } N(0, \sigma^2)$
- Sous H , le modèle est :

$$x_{ij} = \mu + \epsilon_{ij}$$

- Sous A , le modèle est :

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

où α_i est *l'effet de la modalité i du facteur A sur la variable X*

- Pour chaque modèle, on peut produire un estimateur de la variance résiduelle

Paired tests : pourquoi pas ?

Pourquoi ne pas commencer en faisant des tests (z ou t) pour chaque paire d'échantillons ?

- Pour m comparaisons (indépendantes), la probabilité de rejeter au moins un H peut s'écrire : $\alpha_m = 1 - (1 - \alpha)^m$; pour $\alpha = 0.05$:
- 3 tests $\implies$ l'erreur de type I = 0.14
- 5 tests $\implies$ l'erreur de type I = 0.23
- 10 tests $\implies$ l'erreur de type I = 0.4
- 21 tests $\implies$ l'erreur de type I = 0.66

Statistique de test

- Sous H , $SCE_{trts}/(k-1)$ et $SCE_{erreur}/(n-k)$
 $\Rightarrow$ *estimateurs du même paramètre σ^2*
- Donc (sous H), le rapport $\frac{SCE_{trts}/(k-1)}{SCE_{erreur}/(n-k)} \approx 1$
- Sous A , au moins 1 $\alpha_j \neq 0$ et $SCE_{erreur}/(n-k)$ est un unique estimateur de σ^2 ; $SCE_{trts}/(k-1) \gg SCE_{erreur}/(n-k)$
- Donc (sous A), le rapport $\frac{SCE_{trts}/(k-1)}{SCE_{erreur}/(n-k)}$ *très supérieur à 1*
- $\Rightarrow$ Test *unilatéral* dans tous les cas
- $F_{obs} = \frac{SCE_{trts}/(k-1)}{SCE_{erreur}/(n-k)} = CM_{trts}/CM_{erreur}$
- Statistique de test distribuée selon une loi de Fisher à $k-1$ (numérateur) et $n-k$ (dénominateur) degrés de liberté (df = degrees of freedom)

Tableau d'ANOVA

Tableau d'ANOVA

source	df	SC (SS)	CM (MS) (=SC/df)	F	p-valeur
traitements	$k - 1$	SCE_{trts}	$SCE_{trts} / (k - 1)$	CM_{trts} / CM_{erreur}	$P(F_{obs} > F_{k-1, n-k})$
erreur	$n - k$	SCE_{erreur}	$SCE_{erreur} / (n - k) (= \hat{\sigma}^2)$		
total (corr.)	$n - 1$	SCE_{totale}			

■ Sortie d'ordinateur – ANOVA

```
> redcell.aov<-aov(Folate~Group)
> summary(redcell.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	2	15516	7758	3.7113	0.04359 *
Residuals	19	39716	2090		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*** Suppositions ***

- *Indépendance* : Les k échantillons comparés sont indépendants; l'ensemble des n individus est réparti *au hasard* (*randomisation*) entre les k modalités du facteur contrôlé A , n_i individus recevant le traitement i .
- *Homoscédasticité* : Les k populations comparées ont la même variance; le facteur A agit seulement sur la *moyenne* de la variable X et ne change pas sa variance
- *Normalité* : La variable quantitative étudiée suit une loi normale dans les k populations comparées (ou TCL s'applique pour les n_i 'suffisamment grands')
- (*voir vidéo cours 11b pour l'évaluation du modèle*, qui **NE SERA PAS EXAMINÉE**)

Exemple 11a.1

- Un essai clinique est mené pour comparer quelques programmes de perte de poids. Le résultat d'intérêt est la *perte de poids* pendant 8 semaines, mesurée en livres.
- *Trois programmes* sont considérés : un régime *hypocalorique*, un régime *faible en gras*, et un régime *faible en glucides*. Un quatrième groupe est considéré comme un *groupe témoin*. On dit aux participants du quatrième groupe qu'ils participent à une étude sur les comportements sains, la perte de poids n'étant qu'un élément d'intérêt.
- Au total, *20 patients* sont affectés au hasard à l'un des 4 groupes. Les poids sont mesurés au départ et les patients sont conseillés sur la bonne mise en oeuvre du régime alimentaire assigné (à l'exception du groupe témoin). Après 8 semaines, le poids de chaque patient est de nouveau mesuré et la différence de poids est calculée.

Exemple 11a.1, cont.

- Pourquoi inclure un groupe témoin ?

Tableau d'ANOVA

source	df	<i>SC</i>	<i>CM</i>	<i>F</i>	<i>p</i> -valeur
			25		0.0014506
erreur					
total (corr.)		123			

- Quelles sont vos conclusions ?

Qu'est-ce que cela veut dire quand on rejette H ?

- L'hypothèse nulle H est conjointe : que *toutes* les moyennes des populations sont égales
- Lorsqu'on rejette l'hypothèse nulle, cela ne signifie pas que les moyennes sont toutes différentes !
- Cela signifie qu'*au moins une* est différente
- Pour en savoir qui est différente, on peut faire des tests 'post-hoc' / *a posteriori* (paires de t -tests, par exemple)

ANOVA : après le test

- Une fois que toutes les conditions d'une ANOVA ont été vérifiées et que l'analyse a été effectuée, deux conclusions sont possibles :
 - on rejette H
 - on n'a pas assez de preuves pour rejeter H
- Si on ne rejette pas H , on conclut qu'il n'y a pas de différences significatives entre les groupes
- Si on rejette H , on veut *identifier les modalités/niveaux du facteur* qui sont responsables du résultat significatif
- (*voir vidéo cours 11c*)