Tester en une seule procédure l'hypothèse d'égalité de plus de 2 moyennes

Hypothèse nulle H_o : $\mu_1 = \mu_2 = \ldots = \mu_k$

Plus qu'une procédure de test, un langage !

k échantillons comparés,

différences significatives ?

proviennent-ils de la même population d'origine ?

→ **Variance inter-échantillons**

→ **Variance intra-échantillons**

Les k populations sont supposées normales, avec une variance commune σ^2

Mais les conclusions sont approximativement valables même pour des populations non normales

Variance inter-échantillons

Comparaison de 3 machines.

Dans l'espoir que les fluctuations aléatoires se compensent en moyenne,
on prélève sur chaque machine un échantillon aléatoire de la production obtenue au cours de 5 périodes différentes (V.A. : X = 'Volume déposé (en ml) dans flacon de contenance 50 ml ').

	Machine 1	Machine 2	Machine 3	
	47	55	54	
	53	54	50	
	49	58	51	
	50	61	51	
	46	52	49	
$\overline{X}_i$	$\overline{X}_1 = 49$	$\overline{X}_2 = 56$	$\overline{X}_3 = 51$	$\overline{\overline{X}} = 52$

H_o : $\mu_1 = \mu_2 = \mu_3$

Variance inter-échantillons

H_o : $\mu_1 = \mu_2 = \mu_3$ [Les 3 populations sont toutes supposées normales]

Les différences entre les moyennes $\overline{X}_i$ sont-elles assez grandes pour indiquer une différence entre les μ_i sous-jacents?

Pour tester l'hypothèse nulle il faut d'abord déterminer les écarts entre les moyennes d'échantillons $\overline{X}_i$ après le calcul de $\overline{\overline{X}}$, moyenne globale des $\overline{X}_i$. On calcule alors leur variance (en bas à droite du tableau des valeurs).

Si k est le nombre d'échantillons :

$$S^2_{\overline{X}} = \frac{1}{(k-1)} \sum_{i=1}^{k} (\overline{X}_i - \overline{\overline{X}})^2$$

(on reconnaît là l'expression d'une variance pour la variable $\overline{X}$)

On appelle cette variance, la **variance expliquée**

(pouvant provenir de différentes populations si Ho est fausse)

Ici : $S^2_{\overline{X}} = \frac{26}{(3-1)} = 13$

Variance inter-échantillons

Comparaison de 3 machines.

on prélève sur chaque machine un échantillon aléatoire de la production obtenue au cours de 5 périodes différentes (V.A. : X = 'Volume déposé (en ml) dans flacon de contenance 50 ml ').

	Machine 1	Machine 2	Machine 3	
	47	55	54	
	53	54	50	
	49	58	51	
	50	61	51	
	46	52	49	
$\overline{X}_i$	$\overline{X}_1 = 49$	$\overline{X}_2 = 56$	$\overline{X}_3 = 51$	$\overline{\overline{X}} = 52$
$\overline{X}_i - \overline{\overline{X}}$	-3	4	-1	$\Sigma(\overline{X}_i - \overline{\overline{X}}) = 0$
$(\overline{X}_i - \overline{\overline{X}})^2$	9	16	1	$\Sigma(\overline{X}_i - \overline{\overline{X}})^2 = 26$

$H_o : \mu_1 = \mu_2 = \mu_3$

Variance expliquée

$$S^2_{\overline{X}} = \frac{26}{(3-1)} = 13$$

Variance intra-échantillons

La variance entre machines, qui vient d'être calculée, n'est pas la seule source de variation.

Il faut également évaluer les fluctuations aléatoires à l'intérieur de chaque échantillon.

Intuitivement, la mesure de ces fluctuations semble être la dispersion (plus exactement la variance) des valeurs observées au sein de chaque échantillon.

On calcule les carrés des écarts au sein de chaque échantillon en utilisant le tableau :

Pour les n observations (indice j) de l'échantillon 1 ,

$$\sum_{j=1}^{n} (X_{1j} - \overline{X}_1)^2 = (47 - 49)^2 + (53 - 49)^2 + (49 - 49)^2 + (50 - 49)^2 + (46 - 49)^2 = 30$$

On calcule de même les carrés des écarts dans les 2èmes et 3èmes échantillons et on en fait la somme. Puis on divise par le nombre total de degrés de libertés pour l'ensemble des trois échantillons (n-1 = 4 ddl).

On obtient ainsi la variance commune S_p^2 (exactement comme dans le cas de 2 échantillons)

Variance intra-échantillons

On appelle cette variance commune S_p^2 , la **variance inexpliquée** parce qu'elle est la variation aléatoire qui ne peut être expliquée systématiquement (par les différences entre machine)

La généralisation s'obtient en considérant l'expression pour **k** échantillons ayant chacun **n** observations :

$$S_p^2 = \frac{(n-1)\sigma_{o1}^2 + (n-1)\sigma_{o2}^2 + \ldots + (n-1)\sigma_{ok}^2}{k(n-1)}$$

D'où

$$S_p^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2}{k(n-1)}$$

mais aussi : $S_p^2 = \frac{\sigma_{o1}^2 + \ldots + \sigma_{ok}^2}{k}$

Ici : $S_p^2 = \frac{30 + 50 + 14}{(4 + 4 + 4)} = 7.83$

D - Inférence Statistique - Estimation et Tests d'hypothèses

6. ANOVA - Analyse de variance à un et à deux facteurs

Facteur machine	Machine 1	Machine 2	Machine 3				
	47	55	54				
	53	54	50				
	49	58	51				
	50	61	51	k	3		
	46	52	49	n	5		
$\bar{x}_i$	49.00	56.00	51.00	$\bar{\bar{x}}$	52.00		
				S_x^2	13.00	Facteur A	Variance Expliquée inter échantillon
				nS_x^2	65.00		
σ_{oi}	7.5	12.5	3.5	S_p^2	7.83	Résiduelle (erreur)	Variance (commune) inexpliquée intra échantillon
$1/(n-1)\Sigma_j (x_{ij} - \bar{x}_i)^2$							

Le Test F

La question clé est : $S^2_{\bar{x}}$ est-il grand par rapport à S^2_p ?

Autrement dit, le rapport $\dfrac{S^2_{\bar{x}}}{S^2_p}$ est il grand?

Comme la variance des moyennes d'échantillons $\sigma^2_{\bar{x}}$ est d'expression σ^2/n, $\sigma^2 = n\sigma^2_{\bar{x}}$

il faut plutôt estimer S^2_p par $nS^2_{\bar{x}}$

On étudie donc le rapport F de Fisher : $\boxed{F = \dfrac{nS^2_{\bar{x}}}{S^2_p}}$

$$F = \frac{\text{Variance expliquée}}{\text{Variance inexpliquée}}$$

Ce rapport F (variable aléatoire) devant fluctuer autour de 1 sous Ho

Si Ho est fausse (les moyennes μ_k des k populations sont différentes), alors $nS^2_{\bar{x}}$ sera relativement grand par rapport à S^2_p et le rapport F tendra à être bien plus grand que 1.

(le numérateur augmentera parce que la différence entre les moyennes des populations entraînera une grande dispersion des moyennes d'échantillons. Alors que le dénominateur continuera à estimer σ^2)

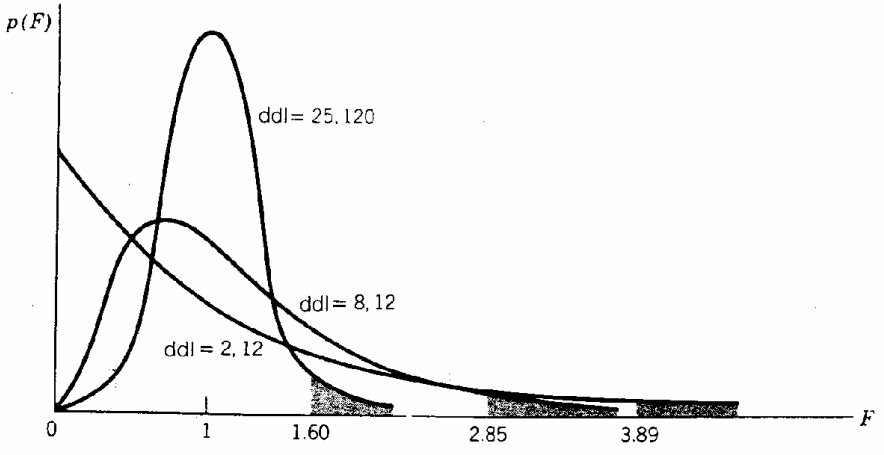
Ainsi plus F est grand, moins l'hypothèse nulle est crédible.

Le Test F

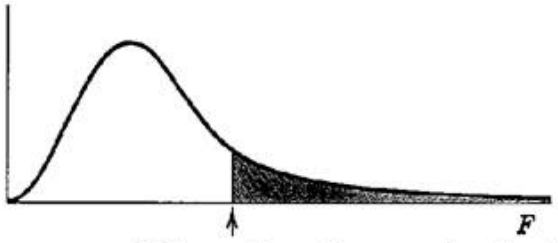
Pour mesurer la crédibilité de Ho, on calcule sa probabilité critique (probabilité de la queue de distribution de F située au delà de la valeur observée).

Lecture de la table de Fisher dépendant des degrés de libertés de la variance du numérateur (k-1) et de ceux de la variance du dénominateur k(n-1)

Table F à double entrée ddl : colonne k-1 et lignes k(n-1)



Quelques distributions particulières de F, avec divers d.d.l. au numérateur et au dénominateur. A noter comment le point critique de 5 % (au-delà duquel H_0 est habituellement rejetée) se déplace vers la gauche, vers la valeur 1, quand d.d.l. augmente.



Valeur critique. Par exemple : $F_{.05}$ signifie qu'on a une probabilité de .05 en queue de distribution.

F = 5x13 / 7.83 = 8.3

ddl :

(3-1) = 2 au numérateur

et 3x(5-1)=12 au dénominateur

8.3 > $F_{0.01}$= 6.93

pour ces ddl

Degrés de liberté pour le dénominateur		Degrés de liberté pour le numérateur: 1	2	3	4	5	6	8	10	20	40	∞
10	$F_{.25}$	1.49	1.60	1.60	1.59	1.59	1.58	1.56	1.55	1.52	1.51	1.48
	$F_{.10}$	3.28	2.92	2.73	2.61	2.52	2.46	2.38	2.32	2.20	2.13	2.06
	$F_{.05}$	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.77	2.66	2.54
	$F_{.01}$	10.0	7.56	6.55	5.99	5.64	5.39	5.06	4.85	4.41	4.17	3.91
	$F_{.001}$	21.0	14.9	12.6	11.3	10.5	9.92	9.20	8.75	7.80	7.30	6.76
12	$F_{.25}$	1.56	1.56	1.56	1.55	1.54	1.53	1.51	1.50	1.47	1.45	1.42
	$F_{.10}$	3.18	2.81	2.61	2.48	2.39	2.33	2.24	2.19	2.06	1.99	1.90
	$F_{.05}$	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.54	2.43	2.30
	$F_{.01}$	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.30	3.86	3.62	3.36
	$F_{.001}$	18.6	13.0	10.8	9.63	8.89	8.38	7.71	7.29	6.40	5.93	5.42
14	$F_{.25}$	1.44	1.53	1.53	1.52	1.51	1.50	1.48	1.46	1.43	1.41	1.38
	$F_{.10}$	3.10	2.73	2.52	2.39	2.31	2.24	2.15	2.10	1.96	1.89	1.80
	$F_{.05}$	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.60	2.39	2.27	2.13
	$F_{.01}$	8.86	5.51	5.56	5.04	4.69	4.46	4.14	3.94	3.51	3.27	3.00
	$F_{.001}$	17.1	11.8	9.73	8.62	7.92	7.43	6.80	6.40	5.56	5.10	4.60
16	$F_{.25}$	1.42	1.51	1.51	1.50	1.48	1.48	1.46	1.45	1.40	1.37	1.34
	$F_{.10}$	3.05	2.67	2.46	2.33	2.24	2.18	2.09	2.03	1.89	1.81	1.72
	$F_{.05}$	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.49	2.28	2.15	2.01
	$F_{.01}$	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.69	3.26	3.02	2.75
	$F_{.001}$	16.1	11.0	9.00	7.94	7.27	6.81	6.19	5.81	4.99	4.54	4.06
20	$F_{.25}$	1.40	1.49	1.48	1.46	1.45	1.44	1.42	1.40	1.36	1.33	1.29
	$F_{.10}$	2.97	2.59	2.38	2.25	2.16	2.09	2.00	1.94	1.79	1.71	1.61
	$F_{.05}$	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.12	1.99	1.84
	$F_{.01}$	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.37	2.94	2.69	2.42
	$F_{.001}$	14.8	9.95	8.10	7.10	6.46	6.02	5.44	5.08	4.29	3.86	3.38
30	$F_{.25}$	1.38	1.45	1.44	1.42	1.41	1.39	1.37	1.35	1.30	1.27	1.23
	$F_{.10}$	2.88	2.49	2.28	2.14	2.05	1.98	1.88	1.82	1.67	1.57	1.46
	$F_{.05}$	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	1.93	1.79	1.62
	$F_{.01}$	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.98	2.55	2.30	2.01
	$F_{.001}$	13.3	8.77	7.05	6.12	5.53	5.12	4.58	4.24	3.49	3.07	2.59
40	$F_{.25}$	1.36	1.44	1.42	1.40	1.39	1.37	1.35	1.33	1.28	1.24	1.19
	$F_{.10}$	2.84	2.44	2.23	2.09	2.00	1.93	1.83	1.76	1.61	1.51	1.38
	$F_{.05}$	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	1.84	1.69	1.51
	$F_{.01}$	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.80	2.37	2.11	1.80
	$F_{.001}$	12.6	8.25	6.60	5.70	5.13	4.73	4.21	3.87	3.15	2.73	2.23
60	$F_{.25}$	1.35	1.42	1.41	1.38	1.37	1.35	1.32	1.30	1.25	1.21	1.15
	$F_{.10}$	2.79	2.39	2.18	2.04	1.95	1.87	1.77	1.71	1.54	1.44	1.29
	$F_{.05}$	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99	1.75	1.59	1.39
	$F_{.01}$	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.63	2.20	1.94	1.60
	$F_{.001}$	12.0	7.76	6.17	5.31	4.76	4.37	3.87	3.54	2.83	2.41	1.89
120	$F_{.25}$	1.34	1.40	1.39	1.37	1.35	1.33	1.30	1.28	1.22	1.18	1.10
	$F_{.10}$	2.75	2.35	2.13	1.99	1.90	1.82	1.72	1.65	1.48	1.37	1.19
	$F_{.05}$	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.91	1.66	1.50	1.25
	$F_{.01}$	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.47	2.03	1.76	1.38
	$F_{.001}$	11.4	7.32	5.79	4.95	4.42	4.04	3.55	3.24	2.53	2.11	1.54
∞	$F_{.25}$	1.32	1.39	1.37	1.35	1.33	1.31	1.28	1.25	1.19	1.14	1.00
	$F_{.10}$	2.71	2.30	2.08	1.94	1.85	1.77	1.67	1.60	1.42	1.30	1.00
	$F_{.05}$	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.83	1.57	1.39	1.00
	$F_{.01}$	6.63	4.61	3.78	3.32	3.02	2.80	2.51	2.32	1.88	1.59	1.00
	$F_{.001}$	10.8	6.91	5.42	4.62	4.10	3.74	3.27	2.96	2.27	1.84	1.00

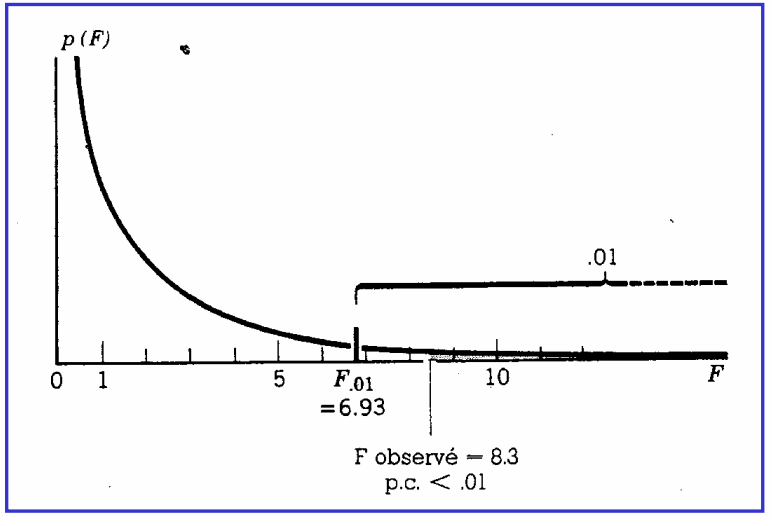
Résolution du test :

F = 5x13 / 7.83 = 8.3

ddl : (3-1) = 2 au numérateur et 3x(5-1)=12 au dénominateur

La table de Fisher révèle que F=8.3 est supérieur à $F_{0.01}$=6.93 pour ces ddl

Cela signifie que, sous Ho, il y a moins de 1% de chances d'obtenir des moyennes d'échantillon qui diffèrent d'autant.



⇨ **On rejette donc Ho**

Conclusion : les productions des 3 machines sont significativement différentes

(avec un risque inférieur à 1% de se tromper en rejetant Ho)

D - Inférence Statistique - Estimation et Tests d'hypothèses

6. ANOVA - Analyse de variance à un et à deux facteurs

	Machine 1	Machine 2	Machine 3		
	47 53 49 50 46	55 54 58 61 52	54 50 51 51 49	$H_o : \mu_1 = \mu_2 = \mu_3$	
$\overline{X}_i$	49,00	56,00	51,00	k n $\overline{\overline{X}}$	**3** **5** **52,00**
$1/(n-1)\Sigma_j (X_{ij} - \overline{X}_i)^2$	7,5	12,5	3,5	S_x^2 S_p^2 F α_o	**13,00** **7,83** **8,30** **0,01**

<u>Conclusion</u> : **Différences significatives au seuil de 5 %**

Le test avec R .

volume	machine
47	1
53	1
49	1
50	1
46	1
55	2
54	2
58	2
61	2
52	2
54	3
50	3
51	3
51	3
49	3

⇨ **étape 1 :**
créer le fichier texte des données
(2 colonnes, 2 titres de colonnes,
ils deviendront 2 variables sous R).

machines1.txt

Le test avec R .

```
> production<-read.table("machines1.txt",h=TRUE)
> summary(production)
     volume          machine
 Min.   :46.0   Min.   :1
 1st Qu.:49.5   1st Qu.:1
 Median :51.0   Median :2
 Mean   :52.0   Mean   :2
 3rd Qu.:54.0   3rd Qu.:3
 Max.   :61.0   Max.   :3
> attach(production)
> machine
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
> summary(volume)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   46.0    49.5    51.0    52.0    54.0    61.0
```

Le test avec R .

```
> facmac<-as.factor(machine)
> anova(aov(volume~facmac))
Analysis of Variance Table

Response: volume
          Df  Sum Sq Mean Sq F value   Pr(>F)
facmac     2 130.000  65.000  8.2979 0.005461 **
Residuals 12  94.000   7.833
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lorsque H_0 est rejetée,

on peut comparer toutes les moyennes 2 à 2 à l'aide d'un test **t de Student** pour en déterminer la cause, il est également utile de procéder à des **combinaisons linéaires de moyennes** permettant la **comparaison de 2 moyennes** à partir d'un ensemble de plus de 2 moyennes (**contraste**).

On rajoute une étape mais c'est indispensable pour conclure.

Exemple de contraste :

Un biochimiste mesure la solubilité d'acides aminés dans un solvant organique.

Il obtient des mesures pour la glycine (solubilité moyenne μ_1), la phénylalanine (μ_2), la tyrosine (μ_3) et le tryptophane (μ_4).

Il est logique de tester **H_0 : $\mu_1 - (\mu_2+\mu_4)/2 = 0$**

Autrement dit, la solubilité d'une petite chaîne latérale comme la glycine est la même que celle d'une grande chaîne latérale contenant un cycle aromatique hydrophobe.

Plusieurs facteurs impliqués dans la même analyse

…. ce sera pour une prochaine fois!