Analyse de la variance

ABDELOUAHAB A

Position du problème :
Comparaison de deux moyennes
→ test de l’écart réduit
Comparaison de plusieurs moyennes ?

Solution :
Comparaison deux à deux :
Inconvénients : méthode longue
  Vue d’ensemble flou
Analyse de la variance : bon reflet de dispersion autour de la moyenne.
Test approprié : test de Fisher-Snedecor(analyse de la variance)

Principe de solution :
Il repose sur la décomposition de la variation de la variable X
1) La variation totale(D totale) de la variable X est mesurée par la dispersion de toutes les valeurs (N valeurs) autour de la moyenne générale
Cette variation dépend de deux sources de variations :
2) La variation entre les classes (variation inter-groupe)
(D inter) Chaque classe est caractérisée par sa moyenne ($\bar{x}_1, ..., \bar{x}_k$) qui s'écarte plus ou moins de la moyenne générale ($\bar{x}$). Effet du facteur étudié : variable indépendante

3) La variation à l'intérieur de chaque classe (intra-groupe, résiduelle) (D intra)

i) La relation entre les variations

\[ D_{totale} = D_{inter} + D_{intra} \]

---

**Principe du test:**
On teste le rapport de deux variances

- variance **inter-groupe** (variance factorielle Vf)
- variance **intra-groupe** (variance résiduelle Vr)

Ho teste donc l'hypothèse de l'homogénéité des k moyennes.
On dit aussi que le facteur (à partir duquel on a construit les k groupes) n'a pas d'influence sur la variable X
Le rapport suit une loi F de Fisher - Snedecor

---

1-variance factorielle Vf

\[ V_f = \frac{D_{inter}}{k-1} \]

K : nombre d'échantillons

**Dispersion inter-groupes** : analyse la variabilité

\[ D_{inter} = n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2 + n_3 (\bar{x}_3 - \bar{x})^2 + ... + n_k (\bar{x}_k - \bar{x})^2 \]

**Formule pratique**

\[ D_{inter} = n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2 + ... + N (\bar{x})^2 \]

$X_i$ : les moyennes respectives des K échantillons

$\bar{X}$ : moyenne générale = $\frac{\Sigma{X_i}}{n_1+n_2+n_3+...+nk}$

N=\(n_1+n_2+n_3+...+nk\)

---

2-variance résiduelle Vr :

\[ V_r = \frac{D_{intra}}{N-k} \]

N= somme des tailles de tous les échantillons :

N=\(n_1+n_2+n_3+...+nk\)

**Dispersion intra-groupes** : analyse la variabilité

\[ D_{intra} = \Sigma{(X_{i1} - \bar{X}_i)^2} + \Sigma{(X_{i2} - \bar{X}_i)^2} + ... + \Sigma{(X_{ik} - \bar{X}_i)^2} \]

**Formule pratique**

\[ D_{intra} = \Sigma{X_{i1}^2} + \Sigma{X_{i2}^2} + ... + nkX_{ik}^2 \]

\[ \Sigma{X_{i1}} = \Sigma{x_{i1}} + \Sigma{x_{i2}} + ... + \Sigma{x_{ik}} \]

\[ \Sigma{x_{i1}^2} = \Sigma{x_{i1}^2} + \Sigma{x_{i2}^2} + ... + \Sigma{x_{ik}^2} \]

2
3- la table de Fisher, donne directement $F$ (valeur critique) selon le risque $\alpha$ choisi

\[ \text{l'intersection de} \quad \text{ddl} = K - 1 \quad \text{ddl} = N - K \]

**Procédure**

1) Détermination du risque $\alpha$
2) Formulation des hypothèses $H_0$ et $H_1$
3) Calcul de la dispersion inter-groupe
4) Calcul de la dispersion intra-groupe
5) Calcul de la variance factorielle $V_f$
6) Calcul de la variance résiduelle $V_r$
7) Conclusion : comparaison du rapport $V_f/V_r$ avec $F$ de la table:

**Conditions d'utilisation**

1- L'indépendance des échantillons.
2- L'homogénéité des variances (c'est-à-dire égalité des variances). Il convient alors de réaliser un test d'homogénéité :
   - test de Bartlett si les effectifs des échantillons sont différents ;
   - test de Cochran si les effectifs sont égaux.
3- Normalité de la distribution des mesures

Le non-respect de ces conditions peut entraîner des résultats erronés soit en acceptant $H_0$ soit en rejetant $H_0$.

Lorsque les variances sont hétérogènes, l'analyse de variance n'est pas utilisable et on recours à un test de Kruskal-Wallis
Exemple :
5 milieux de culture de BCG
10 tubes par milieu de culture
on observe le nombre de colonies par tubes
question : ces milieux sont-ils équivalents?

<table>
<thead>
<tr>
<th>Milieu de culture</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>11</td>
<td>7</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>18</td>
<td>14</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>12</td>
<td>10</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>15</td>
<td>11</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>13</td>
<td>9</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>8</td>
<td>10</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>15</td>
<td>9</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>16</td>
<td>11</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>3</td>
<td>9</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Moyenne</td>
<td>9.5</td>
<td>13.0</td>
<td>9.7</td>
<td>10.5</td>
<td>7.0</td>
</tr>
</tbody>
</table>

1) \( \alpha = 5\% \)
2) \( H_0 \): les milieux sont équivalents
3) \( \text{Dinter} = n1(\bar{x}_1)^2 + n2(\bar{x}_2)^2 + \cdots + N(\bar{x})^2 = 10(9.5)^2 + 10(13)^2 + \cdots + 50(9.94)^2 = 185.7 \)
4) \( \text{Dintra} = \sum x_i^2 - (n1.x_1^2 + n2.x_2^2 + \cdots + nk.x_k^2) = 5083 - (10 \cdot 9.5^2 + 10 \cdot 13^2 + \cdots + 50 \cdot 9.94^2) = 225.1 \)
5) \( V_f = 185.7/5 - 1 = 46.4 \)
6) \( V_r = 225.1/50 - 5 = 5 \)
7) conclusion : \( V_f / V_r = 46.4/5 = 9.3 \)
F de la table = 2.5 (ddl1 = 4 et ddl2 = 45)
le rapport est supérieur à F critique
\( H_0 \) est rejetée.

Comparaison de deux variances
ABDELOUAHAB A
Intérêt :
1) comparer deux moyennes dans le cas des petits échantillons, le test de Student n’est valide qu’au conditions suivantes:
- la variable étudiée doit être normale
- les variance des deux populations ne doivent pas être significativement différentes
2) comparer la précision de deux méthodes

Principe:
Comparaison du rapport $F = \frac{S_A^2}{S_B^2}$ ($S_A^2$ est supérieure à $S_B^2$) et $F$ de la table de Fisher.
F critique: intersection de ddf=nA-1 et ddf=nB-1

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>36</td>
<td>39</td>
</tr>
<tr>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>38</td>
<td>37</td>
</tr>
<tr>
<td>37</td>
<td>36</td>
</tr>
<tr>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>38</td>
<td>37</td>
</tr>
</tbody>
</table>

Moyenne 37 38
Variance 0.5 1.11

Procédure:
Les étapes à suivre sont:
1) Détermination du risque α
2) construction de l’hypothèse nulle
3) Calcul de la variance testée $F = \frac{S_A^2}{S_B^2}$
4) Détermination de la valeur critique $F$ de la table de Fischer, en fonction du risque α et les degrés de liberté $V_1=nA-1$ et $V_2=nB-1$
5) Conclusions :
$F_0$: Si la différence est significative au risque α et $H_0$ est rejetée
$F < F_t$ la différence n’est pas significative au risque α et $H_0$ est retenue
Exemple 2 :
Trois méthodes pédagogiques sont utilisées sur trois groupes d'adultes de même niveau initial. Après l'enseignement les performances sont les suivantes :

<table>
<thead>
<tr>
<th>Enseignement</th>
<th>Effectif</th>
<th>Total</th>
<th>Moyenne</th>
</tr>
</thead>
<tbody>
<tr>
<td>livresque</td>
<td>18</td>
<td>198</td>
<td>11</td>
</tr>
<tr>
<td>audiovisuel</td>
<td>20</td>
<td>240</td>
<td>12</td>
</tr>
<tr>
<td>ordinateur</td>
<td>22</td>
<td>308</td>
<td>14</td>
</tr>
</tbody>
</table>

Sachant que $\sum x^2 = 9967$, peut-on apporter la preuve d'une différence d'efficacité entre ces trois méthodes à $p = 0.01$ ?

<table>
<thead>
<tr>
<th>Source de variation</th>
<th>Somme des carrés des écarts</th>
<th>Nombre de ddl</th>
<th>Carres Moyens (variances)</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>entre les groupes (inter)</td>
<td>94.73</td>
<td><strong>3 - 1 = 2</strong></td>
<td>47.36</td>
<td>4.52</td>
</tr>
<tr>
<td>à l'intérieur des groupes (intra)</td>
<td>597.60</td>
<td><strong>60 - 3 = 57</strong></td>
<td>10.47</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td><strong>691.73</strong></td>
<td><strong>50</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lecture du $F$ de Snedecor : $k - 1 = 2$, $N - k = 57$
pour $p = 0.01$ $F_{lu} = 5.18$ $F$ calculé $< F_{lu}$
donc Acceptation de Ho
La preuve d'une différence ne peut être apportée (à $p = 0.01$).

L'étude de la distribution du taux de cholestérol dans deux groupes de sujets adultes adonné les résultats en g/l:

<table>
<thead>
<tr>
<th>Taille</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Echantillon1</td>
<td>50</td>
</tr>
<tr>
<td>Echantillon2</td>
<td>62</td>
</tr>
</tbody>
</table>

$F$ calculé=1.15
$F$ de la table=1.60(5%, 49, 61)