ii. Transformations de rang

La procédure de Kruskal-Wallis Test est une technique très puissante et utile qui remplace les observations par leurs rangs ; elle est connue-appelée, aussi, méthode de transformation de rang.

D'après Conover (1980), la procédure de Kruskal-Wallis Test est équivalente à l'application de l'analyse de variance usuelle, standard sur les rangs. En effet, en appliquant le Fisher-test standard sur les rangs plutôt que sur les données d'origines, on définit la statistique du test comme suit,

$$F_0 = \frac{H/(a-1)}{(N-1-H)/(N-a)}$$

Notons que si Kruskal − Wallis statistique H augmente ou diminue
alors Fisher statistique standard F_0 augmente ou diminue, aussi.

La transformation de rang est largement applicable dans les problèmes de plans d'expériences où il n'existe pas d'alternative non paramétrique pour l'analyse de variance. Si les données sont rangées et le F-test ordinaire, standard est appliqué, il résulte une procédure approximative de propriétés statistiques parfaites. Aussi, lorsqu'on se doute de l'hypothèse de normalité ou de l'effet des extrêmes ou des valeurs irrégulières, l'analyse de variance usuelle, standard doit être exécutée sur les données d'origines et sur les rangs. Si les deux procédures donnent des résultats similaires, les hypothèses d'analyse de la variance sont probablement et raisonnablement bien satisfaites, et l'analyse standard est correcte et acceptable. Par contre, si les résultats des deux procédures diffèrent, alors la transformation de rang doit être préférée étant donné qu'elle est moins susceptible d'être déformée par des observations non normales et inhabituelles.

Dans de tels cas, l'expérimentateur peut, aussi, étudier l'utilisation des transformations de rang pour la non-normalité, examiner les données et la procédure expérimentale pour déterminer la présence possible des valeurs aberrantes et la cause de leur formation.

Chapitre 4, Blocs aléatoires, Carrés Latin et Plans liés

1- Plan Bloc Complètement Randomisé (RCBD):

Dans toute expérience, la variabilité résultante d'un facteur de nuisance peut affecter les résultats. Généralement, un facteur de nuisance est un facteur de plan qui a un effet sur le résultat-réponse. Parfois, ce facteur de nuisance est négligé; il peut être inconnu par suite incontrôlable au cours de l'expérience. La technique de randomisation du plan d'expérience est utilisée pour se prémunir contre un tel facteur de nuisance caché. Aussi, ce facteur de nuisance peut être connu mais non contrôlable. Si on observe au moins sa valeur qu'il prend à chaque exécution de l'expérience, on peut le corriger dans l'analyse statistique par la technique de l'analyse de la covariance (par exemple). Lorsque la variabilité de source de nuisance est connue et contrôlable, une technique de plan extrêmement importante, appelée Blocking[1] ou contrôle de l'erreur, est appliquée pour éliminer systématiquement l'effet de nuisance sur les comparaisons statistiques possibles entre traitements; technique statistique utilisée considérablement dans les expériences industrielles.

Illustration de l'idée générale : supposant qu'on souhaite déterminer si ou non quatre différentes pointes produisent des lectures différentes sur une machine d'essai de dureté. La machine fonctionne en enfonçant la pointe dans un coupon[2]-test métallique et à partir de la profondeur de l'enfoncement[3] résultant, la dureté du coupon peut être déterminée. Soit, on décide d'obtenir quatre observations pour chaque pointe. Dans l'expérience, il ya un seul facteur, de type pointe, et un plan complètement randomisé à un facteur de contrôle qui consiste à attribuer au hasard chacun des $(4 \times 4 =)16$ essais à une unité expérimentale, qui est coupon métallique, et à observer la mesure de dureté qui en résulte. Ainsi, 16 coupons-test métalliques différents seraient nécessaires dans cette expérience, un pour chaque essai dans le plan.

Cependant, en cette condition de plan et avec une expérience complètement randomisée, on constate un problème potentiellement sérieux. En effet, si les coupons métalliques diffèrent légèrement en leur dureté, s'ils sont prélevés sur des lingots produits à des températures différentes, les unités expérimentales (i.e. coupons-test métalliques) contribueront ainsi à la variabilité observée dans les données-réponses de dureté. Par conséquent, on remarque que l'erreur expérimentale reflètera à la fois l'erreur aléatoire et la variabilité entre coupons.

[1] Formation de Blocs.
[2] Appelé aussi éprouvette métallique.
[3] Ou affaissement.

Cependant, on essaye de rendre l'erreur expérimentale aussi petite que possible en supprimant « si possible » la variabilité entre coupons de cette erreur expérimentale. On adopte, ainsi, un plan expérimental qui teste chaque pointe une fois sur chacun des quatre coupons, voir tableau ; plan expérimental appelé Plan Bloc Complètement Randomisé (RCBD).

Le plan est caractérisé de « complet » étant donné que chaque bloc-coupon contient tous les traitements-pointes. Les blocs-coupons forment des unités expérimentales plus homogènes sur lesquelles on essaye de comparer les traitements-pointes. Par cette stratégie de plan adopté, on assiste à une amélioration de la précision des comparaisons entre les traitements-pointes en éliminant la variabilité entre coupons (ou blocs-coupons). Dans un bloc-coupon, l'ordre dans lequel les quatre (traitements)-pointes sont testées, est déterminé aléatoirement. Notons que ce RCBD plan est une généralisation du concept de plan de comparaisons appariées avec sa procédure du $t - test$ apparié, traité au chapitre 2.

Titre : Plan Bloc Complètement Randomisé (RCBD) pour l'expérience testant la dureté.

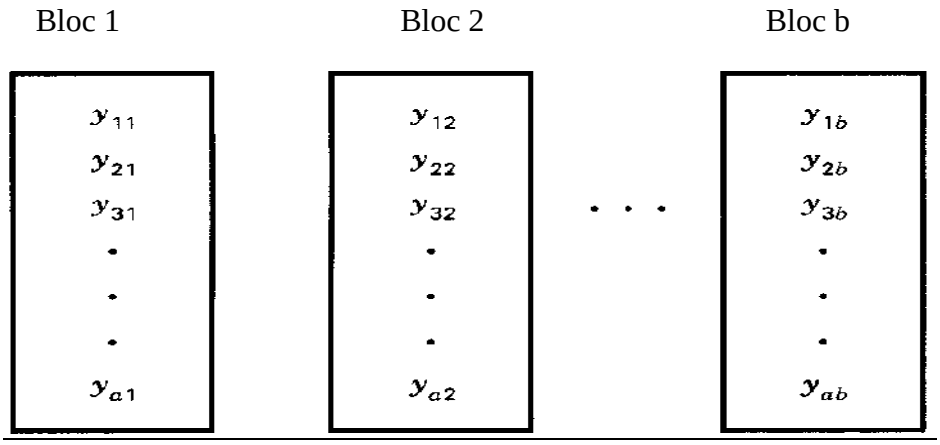
Coupon-test (i.e. Bloc-coupon)

Type de pointe	1	2	3	4
1	9.3	9.4	9.6	10.0
2	9.4	9.3	9.8	9.9
3	9.2	9.4	9.5	9.7
4	9.7	9.6	10.0	10.2

a. Analyse statistique du plan RCBD :

En général, on suppose une expérience de a traitements à comparer et b blocs. Il y a une observation par traitement dans chaque bloc et l'ordre dans lequel les traitements sont testés à l'intérieur de chaque bloc est déterminé aléatoirement. Etant donné que la randomisation des traitements est à l'intérieure des blocs, on dit que les blocs représentent une restriction à la randomisation.

Titre : Plan Bloc Complètement Randomisé (RCBD).



Pour un plan RCBD, le modèle statistique peut être défini comme modèle d'effets, sur-spécifié où

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \qquad (1)$$

$$\mu \, moyenne \, globale; \; \tau_i \, effet \, du \, ième \, traitement; \; \beta_j \, effet \, du \, jème \, bloc$$

$$\varepsilon_{ij} \rightsquigarrow iid(0, \sigma^2), \qquad \forall \; i = 1,2,\cdots,a \quad et \; j = 1,2,\cdots,b$$

On considère qu'initialement les traitements et blocs sont des facteurs fixes. Étant donné que dans un plan RCBD, le modèle d'effets est sur-spécifié, on considère alors que généralement les effets de traitement et de bloc comme des écarts par rapport à la moyenne globale, de sorte que

$$\sum_{i=1}^{a} \tau_i = 0 \qquad et \qquad \sum_{j=1}^{b} \beta_j = 0$$

Aussi, il est possible d'utiliser un modèle de moyennes pour le plan RCBD où

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} \qquad (2)$$

$$o\grave{u} \; \mu_{ij} = \mu + \tau_i + \beta_j$$

$$\forall \; i = 1,2,\cdots,a \quad et \; j = 1,2,\cdots,b$$

Et dans un plan RCBD, on cherche à tester l'égalité des traitements moyens par les d'hypothèses d'intérêt:

$$\begin{cases} H_0 : \mu_1 = \cdots = \mu_a \\ H_a : \mu_i \neq \mu_j \; pour \; au \; moins \; une \; paire \; (i,j) \end{cases}$$

Etant donné que la moyenne du ième traitement est définie par

$$\mu_i = \frac{\sum_{j=1}^{b} (\mu + \tau_i + \beta_j)}{b} = \mu + \tau_i$$

De manière équivalente, on définit les hypothèses d'intérêts antérieures en terme d'effets traitements par

$$\begin{cases} H_0: \tau_1 = \cdots = \tau_a = 0 \\ H_a: \tau_i \neq 0 \ pour \ au \ moins \ un \ i \end{cases}$$

Avec,

- Total de toutes les observations prises par le ième traitement :

$$y_{i.} = \sum_{j=1}^{b} y_{ij} \ , \forall i = 1, \dots, a$$

implique que la moyenne des observations prises par ce ième traitement est

$$\bar{y}_{i.} = \frac{y_{i.}}{b}$$

- Total de toutes les observations dans le bloc j :

$$y_{.j} = \sum_{i=1}^{a} y_{ij} \ , \forall \ j = 1, \dots, b$$

implique que la moyenne des observations dans le bloc j est

$$\bar{y}_{.j} = \frac{y_{.j}}{a}$$

- Total de toutes les observations :

$$y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} = \sum_{i=1}^{a} y_{i.} = \sum_{j=1}^{b} y_{.j} \ où \ N = a\ b \ définit \ le \ nombre \ total \ des \ observations$$

implique que la moyenne de toutes les observations est

$$\bar{y}_{..} = \frac{y_{..}}{N}$$

Ainsi, on définit la somme des carrés corrigée totale par

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}[(\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})]^2$$

En développant à droite l'équation, on définit une partition de la somme totale des carrés:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = b\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + a\sum_{i=1}^{a}(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y}_{..})^2 \qquad (3)$$

$$\Leftrightarrow$$

$$SS_T = SS_{traitements} + SS_{Blocs} + SS_E$$

Remarque : A propos des degrés de liberté associés aux sommes carrées, on a :

- SS_T associé à $ddl = N - 1$ car on a N observations ;
- On a a traitements et b blocs alors $SS_{traitements}\ et\ SS_{Blocs}$ sont associées aux ddls respectifs $a - 1\ et\ b - 1$.
- La somme carrée des résidus est définie comme la différence entre somme carrés entre cellules et sommes carrés de traitements et de blocs. Or, on a ab cellules à $ab - 1$ degré de liberté entre eux ; alors SS_E est associé au degré de liberté suivant :

$$ddl = ab - 1 - (a - 1) - (b - 1) = (a - 1)(b - 1)$$

Fin de remarque.

Et sous l'hypothèse de normalité standard des erreurs, on montre que

$$\frac{SS_{traitements}}{\sigma^2} \rightsquigarrow \mathcal{X}^2(a-1)$$
$$\frac{SS_{blocs}}{\sigma^2} \rightsquigarrow \mathcal{X}^2(b-1)$$
$$\frac{SS_E}{\sigma^2} \rightsquigarrow \mathcal{X}^2[(a-1)(b-1)]$$

Par définition, on montre que si les traitements blocs sont fixes, on a

$$E\left(CM_{SS_{traitements}}\right) = \sigma^2 + \frac{b\sum_{i=1}^{a}\tau_i^{\,2}}{a-1}$$
$$E\left(CM_{SS_{blocs}}\right) = \sigma^2 + \frac{a\sum_{j=1}^{b}\beta_j^{\,2}}{a-1}$$
$$E\left(CM_{SS_E}\right) = \sigma^2$$

Donc, sous l'hypothèse de base d'égalité des traitements moyen, on définit la statistique du test par

$$F_0 = \frac{CM_{SS_{traitements}}}{CM_{SS_E}} \rightsquigarrow F_{[a-1,(a-1)(b-1)]}\ si\ l'hypothèse\ de\ base\ est\ vraie.$$
$$Decision: Rejet\ de\ l'hypotèse\ de\ base\ pour$$
$$F_0 > F_{\alpha,[a-1,(a-1)(b-1)]}$$

Aussi, dans le cas où les moyennes ne diffèrent pas trop, on est amené à comparer les blocs moyens. En tenant compte de la valeur espérée du carré moyen, $CM_{SS_{blocs}}$, on essaye à tester sous l'hypothèse de base, $H_0: \beta_j = 0$, la statistique du test

$$F_0 = \frac{CM_{SS_{Blocs}}}{CM_{SS_E}} \rightsquigarrow F_{[b-1,(a-1)(b-1)]}\ si\ H_0 est\ vraie$$
$$comparée\ à\ la\ valeur\ critique\ F_{\alpha,[b-1,(a-1)(b-1)]}$$

Cependant, on rappelle que l'effet de randomisation (des traitements) est appliqué seulement sur les traitements dans les blocs et signifie que les blocs représentent une restriction à la randomisation, i.e. une restriction d'erreur. A ce niveau, on se demande quel sera son effet sur la statistique du test $F_0 = \frac{CM_{SS_{Blocs}}}{CM_{SS_E}}$???

La réponse à la question a suscité certaines critiques:

- Box-Al. (1978) ont montré que l'analyse de variance standard du F-test peut être justifiée sur la base seulement[4] de la randomisation sans l'utilisation directe de l'hypothèse de normalité. Aussi, ils ont observé que le test de comparaisons des blocs moyens ne peut pas s'imposer avec cette justification en raison de la restriction de randomisation. Aussi, la statistique-test $F_0 = \frac{CM_{SS_{Blocs}}}{CM_{SS_E}}$ ne peut s'appliquer pour la comparaison des blocs moyens que sous la condition résiduelle,

$$\varepsilon_{ij} \rightsquigarrow iid(0, \sigma^2),\ \forall\ i = 1,2,\cdots,a \quad et\ j = 1,2,\cdots,b.$$

- Anderson-Al. (1974) ont appuyé la critique que la restriction de randomisation empêche la statistique-test, $F_0 = \frac{CM_{SS_{Blocs}}}{CM_{SS_E}}$, d'être un test significatif pour comparer les blocs moyens. Ils ont estimé que ce rapport F est vraiment un test pour l'égalité des blocs moyens plus la restriction de randomisation.

En pratique, étant donné que l'hypothèse de normalité est souvent discutable, admettre la statistique $F_0 = \frac{CM_{SS_{Blocs}}}{CM_{SS_E}}$ comme un F-test exact pour tester l'égalité des blocs moyens ne sera pas une bonne utilisation générale. Par conséquent, on exclut ce F-test de la table d'analyse de variance.

Pour enquêter sur l'effet de la variable-bloc, on définit une procédure approximative qui examine raisonnablement le ratio $\frac{CM_{SS_{Blocs}}}{CM_{SS_E}}$. Si le ratio est grand, implique que le facteur-Bloc est d'effet élevé et que la réduction de bruit obtenue par blocage a été

[4] Étant donné que la distribution Fisher en théorie normale est une approximation de la distribution de randomisation générée en calculant la statistique F_0 à partir de chaque affectation possible des réponses aux traitements.

probablement utile pour améliorer la précision de la comparaison des traitements moyens.

Procédure approximative est résumée dans la table d'analyse de variance suivante :

Titre : Analyse de variance pour un plan RCBD

Source De variation	Sommes Carrées	ddl	Carrés Moyesn	StatistiqueF_0
Traitements	$SS_{traitements}$	$a-1$	$\frac{SS_{traitements}}{a-1}$	$\frac{CM_{SS_{traitements}}}{CM_{SS_E}}$
Blocs	SS_{Blocs}	$b-1$	$\frac{SS_{blocs}}{b-1}$	
Erreurs	SS_E	$(a-1)(b-1)$	$\frac{SS_E}{(a-1)(b-1)}$	
Total	SS_T	$N-1$		

Avec :

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SS_{traitements} = \frac{1}{b}\sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{N}$$

$$SS_{Blocs} = \frac{1}{a}\sum_{j=1}^{b} y_{.j}^2 - \frac{y_{..}^2}{N}$$

$$SS_E = SS_T - SS_{traitements} - SS_{Blocs}$$

b. Estimation du modèle statistique et Test de signification de la régression générale:

- Dans le cas où traitements et Blocs sont fixes, on estime le modèle linéaire statistique d'un plan RCBD par approche des Moindres Carrées où :

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \qquad (1)$$
$$\forall\ i = 1,2,\cdots,a \quad et\ j = 1,2,\cdots,b$$
$$défini\ sous\ contraintes\ usuelles: \sum_{i=1}^{a} \tau_i = 0 \qquad et \qquad \sum_{j=1}^{b} \beta_j = 0$$
$$qui\ simplifient\ les\ équations\ normales\ (d'estimation):$$
$$a\,b\hat{\mu} = y_{..}$$
$$b\hat{\mu} + b\hat{\tau}_i = y_{i.}\ \forall\ i = 1,\cdots,a$$
$$a\hat{\mu} + a\hat{\beta}_j = y_{.j}\ \forall\ j = 1,\cdots,b$$
$$\Longrightarrow estimations:$$
$$\hat{\mu} = \overline{y_{..}}$$
$$\hat{\tau}_i = \overline{y_{i.}} - \overline{y_{..}} \quad \forall\ i = 1,\cdots,a$$
$$\hat{\beta}_j = \overline{y_{.J}} - \overline{y_{..}} \quad \forall\ j = 1,\cdots,b$$
$$\Longrightarrow$$
$$\widehat{y_{ij}} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j = \overline{y_{i.}} + \overline{y_{.J}} - \overline{y_{..}}$$
$$et$$
$$pour\ vérification\ de\ l'adéquation\ du\ modèle\ statistique:$$
$$\widehat{\varepsilon_{ij}} = y_{ij} - \widehat{y_{ij}} = y_{ij} - \overline{y_{i.}} - \overline{y_{.J}} + \overline{y_{..}}$$

- Test de signification de la régression générale : procédure aussi utilisée pour développer la procédure d'analyse de variance pour le plan RCBD. En utilisant les solutions-estimations des équations normales, la réduction[5] en somme de carrés pour ajuster le modèle complet est définie par

$$R(\mu,\tau,\beta) = \hat{\mu}\, y_{..} + \sum_{i=1}^{a} \hat{\tau}_i\, y_{i.} + \sum_{j=1}^{b} \hat{\beta}_j\, y_{.j}$$
$$= \sum_{i=1}^{a} \frac{{y_{i.}}^2}{b} + \sum_{j=1}^{b} \frac{{y_{.j}}^2}{a} - \frac{{y_{..}}^2}{ab}$$
$$avec\ (a+b-1)\ degré\ de\ liberté;$$
$$et$$
$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b} \mathrm{y}_{\mathrm{ij}}^2 - R(\mu,\tau,\beta)$$
$$= \sum_{i=1}^{a}\sum_{j=1}^{b} \left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$$
$$de\ degrés\ de\ liberté\ (\mathrm{a}-1)(\mathrm{b}-1).$$

[5] Technique de Réduction de Gauss qui consiste, tant qu'on peut, à éliminer une seule variable à la fois en « complétant les carrés ».

Pour tester l'hypothèse de base $H_0: \tau_i = 0$, on ajuste le modèle contraint (réduit) suivant :

$$y_{ij} = \mu + \beta_j + \varepsilon_{ij}$$

$qui\ définit\ une\ analyse\ de\ variance$ à $un\ facteur\ de\ contrôle.$

$La\ réduction\ en\ somme\ de\ carrés\ pour\ ajuster\ le\ modèle\ contraint\ donne$:

$$R(\mu, \beta) = \sum_{j=1}^{b} \frac{{y_{.j}}^2}{a}\ \ de\ degrés\ de\ liberté\ b.$$

$Donc, la\ somme\ de\ carrés\ du\ coefficient\ \tau_i\ après\ ajustement\ de\ \mu\ et\ \beta_j\ est$

$$R(\tau/\mu, \beta) = R(\mu, \tau, \beta) - R(\mu, \beta)$$

$$= R(modèle\ complet) - R(modèle\ contraint)$$

$$= \sum_{i=1}^{a} \frac{{y_{i.}}^2}{b} + \sum_{j=1}^{b} \frac{{y_{.j}}^2}{a} - \frac{{y_{..}}^2}{ab} - \sum_{j=1}^{b} \frac{{y_{.j}}^2}{a}$$

$$= \sum_{i=1}^{a} \frac{{y_{i.}}^2}{b} - \frac{{y_{..}}^2}{ab} = \mathrm{SS}_{\text{traitements}}$$

à (a − 1) degrés de liberté.

Aussi, la somme carrés Blocs est obtenue par ajustement du modèle contraint (réduit) :

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

$qui\ définit\ aussi\ une\ analyse\ de\ variance$ à $un\ facteur\ de\ contrôle.$

$La\ réduction\ en\ somme\ de\ carrés\ pour\ ajuster\ le\ modèle\ contraint\ donne$:

$$R(\mu, \tau) = \sum_{i=1}^{a} \frac{{y_{i.}}^2}{b}\ \ de\ degrés\ de\ liberté\ \ a$$

$Donc, la\ somme\ de\ carrés\ pour\ Blocs, \beta_j, après\ ajustement\ de\ \mu\ et\ \tau_i, est\ définie$

$$R(\beta/\mu, \tau) = R(\mu, \tau, \beta) - R(\mu, \tau)$$

$$= \sum_{i=1}^{a} \frac{{y_{i.}}^2}{b} + \sum_{j=1}^{b} \frac{{y_{.j}}^2}{a} - \frac{{y_{..}}^2}{ab} - \sum_{i=1}^{a} \frac{{y_{i.}}^2}{b}$$

$$= \sum_{j=1}^{b} \frac{{y_{.j}}^2}{a} - \frac{{y_{..}}^2}{ab} = \mathrm{SS}_{\text{Blocs}}$$

à (b − 1) degrés de liberté.

Notons que la procédure-Test de signification de la régression générale a développé les sommes (de) carrées de Traitements, de Blocs et de l'Erreur pour un plan RCBD. Aussi, remarquons qu'on ne recourt pas à l'utilisation habituelle de cette procédure pour réellement analyser les données dans un bloc complet randomisé ; elle s'avère parfois utile dans des conceptions de blocs randomisés plus générales.