SMART

Système Méthodologique
d'Aide à la Réalisation de Tests

Statistiques de groupe

et analyse des questions de votre épreuve

IFRES LIÈGE université

Une unité de soutien de l'IFRES • Université de Liège

L'analyse des statistiques de groupe

Lorsque la correction d'une épreuve est confiée au SMART, des statistiques de groupe et des questions sont systématiquement fournies. Celles-ci renvoient à différents concepts mathématiques que nous allons brièvement expliquer dans les lignes qui vont suivre. Ces explications sont en partie tirées des notes de cours du professeur F. Pérée.

Les indices de position

La moyenne

La moyenne d'un ensemble de données est l'indice le plus utilisé pour indiquer la tendance centrale de ces données. Cette mesure statistique exprime la **grandeur qu'aurait chacun des membres de l'ensemble s'ils étaient tous identiques sans changer la dimension globale** de l'ensemble.

Techniquement, il s'agit simplement de la somme des données rapportée à leur nombre et dont la formule est la suivante :

$$\text{moyenne} = \frac{\sum \text{score de l'évalué}_i}{\text{nombre total d'évalués}}$$

Exemple : *quatre étudiants avec des scores de 3, 5, 12, et 5*
moyenne = (3 + 5 + 12 + 5) ÷ 4 = **6,25**

La médiane

La médiane est la valeur se situant « *au milieu de la distribution* » lorsque les données sont ordonnées par ordre croissant ou décroissant. Plus simplement encore, la médiane des scores de l'ensemble des étudiants est **le score en-dessous duquel se trouvent 50% des scores et au-dessus duquel se situent 50% des scores**.

Exemple : *quatre étudiants avec des scores de 3, 5, 12, et 5*
médiane = **4**

Les indices de dispersion

La variance

Même si, nous l'avons vu ci-dessus, la moyenne d'une variable se situe « *au centre* » de l'ensemble de ses valeurs, deux variables de même moyenne peuvent tout de même différer considérablement quant à la dispersion de leurs valeurs.

La variance correspond alors à la **dispersion des scores, des notes, par rapport à la moyenne du groupe**. Elle est égale à la « *moyenne des carrés des écarts des valeurs observés à leur moyenne* ».

$$\text{variance} = \frac{\sum (\text{score de l'évalué}_i - \text{moyenne des scores})^2}{\text{nombre total d'évalués}}$$

Exemple : *quatre étudiants avec des scores de 3, 5, 12, et 5*

$$\textbf{variance} = \frac{(3 - 6{,}25)^2 + (5 - 6{,}25)^2 + (12 - 6{,}25)^2 + (5 - 6{,}25)^2}{4} = \textbf{11,6875}$$

Plus la variance est élevée, plus les écarts à la moyenne sont élevés, plus les données sont dispersées.

La variance ne peut être que positive ou nulle (cela provient du fait que l'on élève les écarts à la moyenne au carré).

La variance n'a pas de limite supérieure, sa valeur dépend de l'unité de mesure (et donc de l'ordre de grandeur) des scores et elle s'exprime dans le carré de l'unité de mesure de la variable de départ, ce qui la rend assez difficile à interpréter dans l'absolu. C'est pourquoi il est parfois **plus aisé d'interpréter l'écart-type** qui exprime la dispersion dans le même système d'unités que la moyenne.

L'écart-type

Il s'agit simplement de la **racine carrée** (positive) **de la variance**.

Exemple : *quatre étudiants avec des scores de 3, 5, 12, et 5*
écart-type = √11,6875 = **3,4187**

Un écart-type est toujours positif ou nul. Un écart-type nul signifierait que les scores n'ont aucune dispersion, autrement dit qu'ils sont tous rassemblés sur la valeur centrale, en l'occurrence la moyenne. Dans ce cas, tous les étudiants auraient tous les mêmes scores, qui seraient évidemment par la force des choses les scores moyens. **Plus l'écart-type est élevé, plus les données sont dispersées**.

Le minimum et le maximum

Il s'agit de la valeur la plus petite et de la valeur la plus grande observées dans l'échantillon.

Exemple : *quatre étudiants avec des scores de 3, 5, 12, et 5*
minimum = **3**
maximum = **12**

L'étendue

Il s'agit de la distance entre la plus petite et la plus grande des valeurs observées. En d'autres termes, il s'agit de la différence entre le maximum et le minimum.

Exemple : *quatre étudiants avec des scores de 3, 5, 12, et 5*
étendue = 12 - 3 = **9**

Indice de consistance interne

L'alpha de Cronbach

L'alpha de Cronbach est un indice de fidélité calculé pour l'ensemble du test.

Qu'entend-on par fidélité d'un test ?

Lorsque l'on construit un test, c'est généralement dans le but d'obtenir la mesure la plus fiable possible d'une performance donnée. Si on considère le test comme un instrument de mesure, on espère que cet instrument va mesurer le plus fidèlement possible la performance vraie, c'est-à-dire qu'il va fournir une estimation la plus proche possible du score *vrai* de l'étudiant.

Si tel est bien le cas, tout comme on peut l'attendre d'un mètre qui mesurerait bien la taille réelle d'un objet, on peut s'attendre à ce que le test fournisse la même mesure d'un essai à l'autre, c'est-à-dire d'une passation à l'autre (dans d'autres circonstances, à un autre moment, corrigé par un autre évaluateur,...) mais aussi qu'il fournisse les mêmes résultats qu'un autre test ou qu'une forme parallèle du test censé évaluer la même performance (tout comme deux mètres fournissent la même mesure).

Cependant, cette situation idéale n'est jamais rencontrée dans la réalité car trois sources d'erreur affectent la fidélité :

1. l'instrument lui-même (les questions du test),
2. le contexte de passation,
3. l'individu.

Si le score vrai que l'on cherche à mesurer est une constante, les erreurs qui affectent le score observé, elles, sont variables.

Ces sources d'erreur font que, si nous adressons deux versions parallèles d'un test aux mêmes étudiants, et malgré de multiples précautions, des différences apparaîtront entre les deux séries de scores des étudiants. La corrélation entre les résultats, qui en théorie devrait être parfaite, ne l'est pas en pratique.

Qu'en est-il exactement de l'alpha de Cronbach ?

L'alpha de Cronbach est une mesure de la fidélité qui permet de rendre compte de la première source d'erreur qui affecte les mesures : l'instrument, c'est-à-dire le test lui-même.

L'alpha de Cronbach estime la fidélité du test en utilisant les informations fournies en une seule passation. Il s'agit d'une méthode d'estimation basée sur l'évaluation de la cohérence interne du test.

Pour ce faire, on sépare le test en deux moitiés, on calcule la corrélation entre les deux séries de résultats, on obtient alors ce qu'on appelle le coefficient de bipartition ou « *split-half* ». L'alpha de Cronbach correspond en fait à la moyenne de tous les coefficients de bipartition possibles pour un même test (que l'on aurait obtenus par « *split-half* » successifs en examinant toutes les combinaisons possibles de questions en deux moitiés). On calcule l'alpha de Cronbach avec la formule suivante :

$$\alpha = \frac{nq}{nq - 1}\left[1 - \frac{\sum s_q^2}{s_t^2}\right]$$

- nq = le nombre de questions du test
- s_q^2 = la variance des scores de la question q
- s_t^2 = la variance de la somme de toutes les questions du test

Sa valeur s'établit entre 0 et 1, étant considérée comme « acceptable » à partir de 0,70. Il permet donc l'estimation de la fidélité du score à un test. Par exemple :

- Un alpha de Cronbach de 0,70 signifie que 70% de la variation des scores observés est due aux scores vrais et non aux fluctuations aléatoires.
- Un alpha de Cronbach nul correspondrait au cas où il n'y a aucune part de score vrai dans les scores observés (le test n'a aucune fidélité).
- Un alpha de Cronbach de 1 correspondrait au cas où la part de score vrai dans les scores observés serait maximale et où la prédiction du score vrai à partir du score observé serait parfaite (toutes les questions sont parfaitement fiables, la cohérence interne du test est parfaite).

L'analyse des questions

En plus des statistiques de groupe, une analyse des questions est également réalisée. Chaque question fait l'objet d'une évaluation statistique comportant notamment la mesure du **degré de difficulté** (% de réponses correctes) et du **degré de discrimination** (*coefficient de corrélation point bisériale* ou r.bis).

Après la lecture des formuLOMs de réponse des étudiants, les réponses sont traitées par notre logiciel, qui calcule trois indices pour chaque proposition de chaque question :

- Le pourcentage d'étudiants qui ont choisi la proposition
- Le *coefficient de corrélation point bisériale* (r.bis)
- La certitude moyenne (si utilisation des degrés de certitude)

	Sol. 0	Sol. 1	Sol. 2	Sol. 3	Sol. 4
Q. 42	0,9	**47,7**	39,3	1,9	10,3
r.bis	-0,51	**0,33**	-0,08	-0,16	-0,18
C_{moy}	65,36	**68,48**	69,35	57,50	51,59

Qu'indique le r.bis ?

Le r.bis, ou *coefficient de corrélation point bisériale*, est un indice statistique édumétrique calculé pour chacune des propositions de chaque question et d'une corrélation linéaire entre le score global au test (variable *métrique*) et le choix pour chacune des propositions (variable *dichotomique* : choisie / pas choisie). En d'autres termes, cette statistique permet de vérifier si, en tendance, **les étudiants les meilleurs au test ont choisi la réponse correcte alors que ça ne serait pas le cas des étudiants les plus faibles**. À l'inverse, les *distracteurs* (réponses incorrectes) devraient présenter des taux de choix inférieurs pour les étudiants les plus forts au test, contrairement aux étudiants les plus faibles.

Comment varie-t-il ?

Le r.bis d'une proposition varie entre -1 et +1. Il est positif si la proposition est choisie, en moyenne, par les étudiants qui obtiennent un score total plus élevé au test et d'autant plus grand que la proposition est massivement choisie par les « *meilleurs* ». Un coefficient négatif correspond à la situation opposée.

Lorsqu'un QCM « *fonctionne* » bien, **on s'attend donc à un r.bis positif et suffisamment élevé pour la réponse correcte** et des r.bis négatifs ou proches de zéro pour les autres propositions.

Le r.bis de la réponse correcte peut être considéré comme satisfaisant s'il est supérieur à un seuil qui est fonction du nombre de questions de l'épreuve : il s'agit de l'inverse de la racine carrée du nombre de questions (n), soit 1/√n. Cela signifie que moins il y a de questions, plus le seuil que le r.bis de la réponse correcte devrait dépasser est élevé. Cette valeur est indiquée dans le coin supérieur droit du document r.bis envoyé par le SMART.

Comment l'interpréter et l'utiliser ?

Le r.bis pour une proposition particulière fournit donc de précieuses informations sur **l'adéquation des distracteurs** des questions à choix multiple et permet d'apprécier dans quelle mesure ils sont ambigus ou discriminants.

Quand il s'agit de la proposition correcte, le r.bis permet de **vérifier si la question est réussie**, en moyenne, **par les étudiants qui ont un score global élevé au test**, autrement dit, si ce sont les étudiants bien préparés qui ont répondu correctement et donc, **de voir si la question est discriminante**. Il renseigne aussi sur la cohérence de la question avec le reste du test.

L'examen du r.bis permet donc :

- de **détecter une incohérence** éventuelle entre le résultat à une question donnée et l'ensemble du test,
- d'**analyser la qualité des solutions proposées**.

Associé aux pourcentages de choix pour chaque proposition, il apporte une aide précieuse à la détection d'un problème de correction (proposition erronément renseignée comme correcte) et/ou à la décision éventuelle de supprimer ou valider pour tout le monde une question non discriminante ou encore de rallier un distracteur choisi par les meilleurs étudiants (pour des raisons qui, après analyse, s'avèreraient pertinentes) au *pool* des réponses correctes.

version : 2017-05-17

Photographie pp. 7 :
© Michel Houet, TILT-ULg

SMART — Système Méthodologique d'Aide à la Réalisation de Tests

Quartier Urbanistes 1
Traverse des Architectes, 5B
B-4000 Liège (Sart Tilman)

smart.ulg.ac.be

+32 4 366 2078

smart@ulg.ac.be

Une unité de soutien de l'IFRES • Université de Liège