

## Concept “Différence significative”

Afin de pouvoir expliquer la signification, nous partons d'une population, notée par  $P$ , une variable sociodémographique (ou “sociodémo” en bref) qui est composée d'une série de classes,  $K_{i| 1 \leq i \leq v}$ , et un ICP choisi à l'avance. A l'aide d'une analyse de signification, nous tentons de déterminer pour chaque classe de la sociodémo s'il existe une différence moyenne significative dans les scores d'ICP entre cette classe et toutes les autres classes combinées.

La notion “différence significative” doit être comprise dans le cadre des statistiques. Nous prenons comme exemple générique la population des travailleurs d'une entreprise déterminée et comme sociodémo “profession” comprenant des classes d'ouvriers, d'employés et de cadres. Lors d'une analyse de signification pour un ICP déterminé, nous essayons de savoir en quelle mesure le score de l'ICP pour ouvriers diffère en moyenne de celui des employés et des cadres, en quelle mesure le score ICP pour employés diffère en moyenne du score des ouvriers et des cadres et finalement en quelle mesure le score des cadres dévie en moyenne de celui des ouvriers et des employés.

A chaque étape de l'analyse, nous devons donc déterminer en quelle mesure les scores ICP des deux sous-groupes séparés de la population (le premier représente la classe ICP choisie à l'avance et le deuxième les autres classes combinées) diffèrent en moyenne. Ces groupes sont  $G_1$  et  $G_2$ . Etant donné qu'en statistiques, nous travaillons toujours avec des chiffres exactes, nous utilisons des moyennes de la population  $\mu_{G_1}$  et  $\mu_{G_2}$  ou des médianes de la population  $\eta_{G_1}$  et  $\eta_{G_2}$  afin de déterminer le score moyen par groupe. Une moyenne de la population  $\mu_{G_1}$  ou  $\mu_{G_2}$  représente le score moyen d'un ICP pour tous les éléments de la population qui appartiennent au groupe  $G_1$  ou  $G_2$ . Une médiane de la population  $\eta_{G_1}$  ou  $\eta_{G_2}$  est la valeur inférieure à la moitié des éléments du groupe  $G_1$  ou  $G_2$ . A chaque étape de l'analyse, nous devons donc déterminer pour chaque classe d'ICP en quelle mesure  $\mu_{G_1}$  et  $\mu_{G_2}$  ou  $\eta_{G_1}$  et  $\eta_{G_2}$  diffèrent les uns des autres. (La taille des groupes  $G_1$  et  $G_2$  déterminera si la différence entre  $\mu_{G_1}$  et  $\mu_{G_2}$  ou  $\eta_{G_1}$  et  $\eta_{G_2}$  sera prise en compte).

En statistiques, les valeurs calculées à partir d'une population sont aussi appelées des paramètres de population. En général, dans la pratique, nous ne connaissons pas exactement ces paramètres de population. En effet, une population est souvent décrite à l'aide d'une terminologie conceptuelle comme par exemple “la population active belge” ou contient trop d'éléments qui ne peuvent dès lors pas tous être pris en compte comme par exemple “tous les habitants belges”. En statistiques, afin de pouvoir procéder à des estimations pour des paramètres de population, nous partons d'un échantillon. Il s'agit d'un ensemble d'objets sélectionnés par hasard dans la population représentative au sein de cette population. En général, nous attribuons le symbole  $S_p$  à un échantillon qui est représentatif pour une population  $P$ .

Dans l'exemple générique, l'échantillon contiendrait un nombre de travailleurs sélectionnés par hasard dans l'entreprise. Sur base de cet échantillon, nous pouvons faire des estimations des valeurs des paramètres de population. Une telle estimation est également appelée une statistique. Il est cependant d'importance cruciale qu'une telle statistique contienne une fonction de répartition associée. Il s'agit d'une fonction mathématique qui associe à chaque chiffre la probabilité que la statistique adopte cette valeur.

Pour les calculs pratiques, le *ZebraZone* toolset part donc d'un échantillon  $S_p$  qui est représentatif pour la population  $P$ . Une estimation de  $\mu_{G_1}$  sera calculée comme la somme de tous les scores ICP des éléments de l'échantillon appartenant au groupe  $G_1$  divisé par le nombre total d'éléments de l'échantillon appartenant à  $G_1$ . Nous appelons cela la moyenne de l'échantillon de l'ICP du sous-groupe  $G_1$  de  $S_p$  et lui attribuons le symbole  $m_{G_1}$ . De manière analogue,  $m_{G_2}$  représente la moyenne de l'échantillon du groupe  $G_2$ .

Pour réaliser une estimation de  $\eta_{G_1}$ , les éléments de  $S_p$  qui appartiennent à  $G_1$  sont d'abord classés du plus petit au plus grand. Si  $n_1$  est impair, l'estimation correspondra à l'élément  $(n_1 + 1)/2$  de cette série classée. Si  $n_1$  est un nombre pair, nous utilisons la moyenne de l'élément  $n_1/2$  et  $(n_1 + 2)/2$  de cette série classée comme estimation de la médiane de la population. Ce chiffre est également appelé médiane de l'échantillon des éléments de  $S$  appartenant à  $G_1$  et nous lui attribuons le symbole  $med_{G_1}$ . De manière analogue,  $med_{G_2}$  représente l'estimation de  $\eta_{G_2}$ . Dans notre exemple générique,  $m_{Ouvriers}$  représente la moyenne de l'échantillon des scores ICP de tous les ouvriers et  $med_{Ouvriers}$  la médiane des scores ICP de tous les ouvriers de l'échantillon.

En général, outre leur fonction d'estimation, les statistiques sont également utilisées pour comparer des paramètres de population. Le calcul de signification du *ZebraZone Toolset* suit pleinement cette philosophie. Si  $\mu_{G_1}$  est parfaitement égal à  $\mu_{G_2}$ , il n'est pas certain que  $m_{G_1}$  et  $m_{G_2}$  adoptent exactement les mêmes valeurs. D'ailleurs, les moyennes d'échantillon ne sont que des approximations des moyennes de la population. Il est évident qu'on s'attend à ce que la différence entre  $m_{G_1}$  et  $m_{G_2}$  soit relativement petite lorsque les moyennes de la population sont égales. Cette même logique est d'application pour les médianes de la population  $\eta_{G_1}$  et  $\eta_{G_2}$  avec les estimations  $med_{G_1}$  et  $med_{G_2}$ . Afin de déterminer à partir de quand la différence mesurée entre  $m_{G_1}$  et  $m_{G_2}$  ou  $med_{G_1}$  et  $med_{G_2}$  nous permet de conclure que les paramètres de population  $\mu_{G_1}$  et  $\mu_{G_2}$  ou  $\eta_{G_1}$  et  $\eta_{G_2}$  sont égaux, nous pouvons utiliser des procédures statistiques. Une telle procédure est également appelée un test statistique.

Etant donné que nous nous basons sur des estimations des paramètres de population, une telle heuristique n'est pas conclusive à 100%. Bien que  $\mu_{G_1}$  et  $\mu_{G_2}$  soient égaux, il existe notamment la possibilité que ceci ne soit pas reflété dans l'échantillon mesuré et qu'on en déduise donc à tort que  $\mu_{G_1}$  et  $\mu_{G_2}$  sont différents. Une telle erreur est appelée tout simplement une erreur de type I. Par ailleurs, il se pourrait également qu'une différence entre  $\mu_{G_1}$  et  $\mu_{G_2}$  ne soit pas reflétée dans le test. Déduire que les moyennes de la population sont égales supposerait une erreur de type II. La probabilité de commettre une erreur de type I lors d'un test est appelé seuil de signification ( $\alpha$ ). La probabilité de faire une erreur de type II portera le symbole  $\beta$ . Le chiffre  $(1-\beta)$  représente donc la probabilité que l'on retrouve également la différence entre  $\mu_{G_1}$  et  $\mu_{G_2}$  ou  $\eta_{G_1}$  et  $\eta_{G_2}$  dans le test. Dans la théorie statistique, ceci est indiqué comme la puissance du test.

Dans pratiquement tous les cas, le seuil de signification du test peut être déterminé à l'avance mais pas la puissance. Cependant, le seuil de puissance et de signification ne sont pas indépendants l'un de l'autre. Il semblerait notamment qu'au plus le seuil de signification est élevé (valeur  $\alpha$ -inférieur), au moins la Puissance du test sera élevé. Donc, un test avec un seuil de signification de 0.05 aura en général moins de puissance qu'un test avec une valeur  $\alpha$  de 0.03 par exemple.

On doit donc choisir  $\alpha$  afin d'obtenir une puissance suffisamment grande. Dans la pratique, la valeur 0.05 est souvent proposée comme seuil de signification. Les analyses du *ZebraZone Toolset* sont conformes à cette directive.

Dans l'analyse de signification, outre le choix du seuil de signification, le choix du type de test jouera également un rôle crucial. Nous optons pour le test t paramétrique ou le test de la somme des rangs de Wilcoxon ce qui est un test non paramétrique. Le test t s'appuie sur une série de conditions de la fonction de répartition des moyennes d'échantillon alors que le test de Wilcoxon ne pose pas de telles exigences. Le choix du test en fonction de la situation est principalement déterminé à l'aide de ces conditions et de la puissance du test.

□ **Premier cas: les deux groupes  $G_1$  et  $G_2$  contiennent plus de 30 éléments.**

Dans ce cas-ci, le "théorème limite central"<sup>1</sup> est d'application: il sous-entend que la fonction de répartition de  $m_{G_1}$  ainsi que de  $m_{G_2}$  suivent la courbe en cloche de Gauss<sup>2</sup>. Il s'agit précisément là d'une des conditions cruciales pour effectuer un test t. De plus, dans ce cas-ci, le test t supposera une puissance plus grande qu'avec le test de la somme des rangs de Wilcoxon. C'est donc la raison pour laquelle nous avons choisi le test t avec un seuil de signification de 0.05 ou de 5%. De manière globale, ce test permet de déterminer en quelle mesure les moyennes de la population des groupes  $G_1$  et  $G_2$  diffèrent les unes des autres. À l'aide de  $m_{G_1}$  et de  $m_{G_2}$  et des données de l'échantillon nous calculons un chiffre t et un intervalle I. Si t appartient à cette intervalle, nous en concluons que les moyennes de la population  $\mu_{G_1}$  et  $\mu_{G_2}$  sont égales. Si ce n'est pas le cas, nous en concluons le contraire.

□ **Deuxième cas: un des deux groupes  $G_1$  et/ou  $G_2$  contient moins de 30 éléments.**

Dans ce cas-ci, le "théorème limite central" n'est plus d'application et le test t n'est dès lors plus fiable. Nous choisissons d'utiliser dans ce cas-ci le test de la somme de rangs de Wilcoxon qui ne dépend pas de conditions de la fonction de répartition des moyennes de l'échantillon. Le test de Wilcoxon permet de déterminer à quel point les médianes de la population  $\eta_{G_1}$  et  $\eta_{G_2}$  diffèrent. À l'aide de données de l'échantillon, nous calculons un chiffre W et un intervalle I. Si W appartient à l'intervalle I, nous concluons que les médianes  $\eta_{G_1}$  et  $\eta_{G_2}$  sont égales et dans les autres cas,  $\eta_{G_1}$  et  $\eta_{G_2}$  seront différents. Le seuil de signification sera fixé à 0.05 ou 5%.

---

<sup>1</sup> Une des thèses fondamentales en statistiques. Globalement, la thèse dit que pour un échantillon S de taille suffisamment grande, la fonction de répartition de la moyenne de l'échantillon est précisément la courbe de Gauss.

<sup>2</sup> La courbe de Gauss est la fonction de répartition la plus importante que l'on retrouve très souvent dans la nature. En fait, elle est la traduction mathématique du phénomène de la moyenne. De cette manière il y a par exemple beaucoup de gens qui ont un poids moyen, un nombre moins important à un poids peu élevé ou élevé et quelques exceptions qui ont un poids extrêmement bas ou élevé.