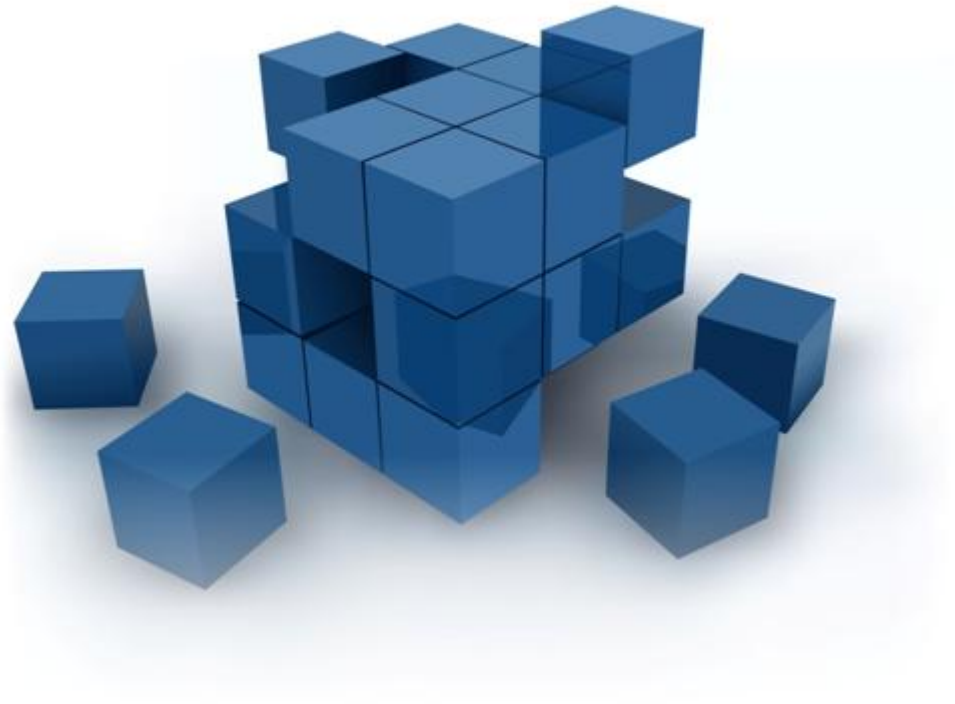


Business Intelligence

Failles et manipulations dans l'analyse de l'information

Un essai de Quentin BRAHAM

Table des matières

Introduction

Avant d'entrer dans le vif du sujet, il est nécessaire d'en présenter tout d'abord le contexte. Ce qui est communément appelé BI (Business Intelligence) se définit par la prise de décisions stratégiques par une organisation, prise de décisions basée sur des informations provenant de systèmes informatiques.

Généralement, ces systèmes informatiques existent au sein même de cette organisation. Ils accumulent toute une série d'informations lors de traitements automatisés, ou d'enregistrements manuels par encodage. Ces informations sont à l'état brut et sont difficilement compréhensibles et exploitables telles quelles par leur format. « L'homme de technique » intervient dès lors pour extraire et reconstruire ces données sous une forme compréhensible, avant de les présenter à « l'homme d'affaires » qui, sur base d'analyse et d'interprétation, pourra prendre les décisions stratégiques nécessaires.

Pour prendre un exemple pratique, imaginons une société qui souhaite développer un produit à l'attention des jeunes. Via les fiches clients enregistrées dans leurs bases de données, elle peut facilement étudier le profil type de leur cible – par leurs habitudes d'achats, par exemple – et développer un produit adapté en conséquence. C'est un exemple simple d'application du processus BI.

Ce processus BI englobe donc deux métiers différents :

- Le premier, informatique, est responsable de la collecte des données mais aussi de l'extraction, du réassemblage, de l'agrégation, de la présentation et de l'accessibilité de l'information. Ce travail peut s'apparenter à celui du statisticien qui, sur base d'une enquête, collecte et trie un ensemble de données, puis en calcule différentes statistiques, avant d'en présenter des conclusions.
- Le second est lié aux affaires, au business, à une stratégie économique. Il est responsable de l'interprétation des conclusions présentées par le premier. Il prend les décisions nécessaires pour atteindre ses objectifs professionnels et améliorer la performance de son entreprise.

Ces informations dominent – que ce soit sous leur forme présentée ou interprétée – le processus décisionnel BI tout au long de son cycle de vie. A tout moment, ces informations peuvent être mal appréhendées et présenter un risque important d'être mal

interprétées, et cela peut influencer, voir manipuler, le processus décisionnel avec des répercussions désastreuses.

Après une première partie consacrée à la présentation de l'information, où l'on observe les principaux outils de présentation, leur utilité et les failles qu'ils apportent dans l'analyse de l'information, la deuxième partie sur l'interprétation de l'information permet d'appréhender quelques caractéristiques importantes de l’information pouvant manipuler le raisonnement et l’interprétation qui en découle. Ce document tend ainsi à démontrer qu'il est important de tenir compte de ces quelques failles énoncées afin premièrement d'effectuer une analyse complète de l'information, et deuxièmement d'éviter toutes manipulations pouvant fausser l'interprétation de cette même information. Et cela pour garantir un processus décisionnel pertinent et sensé.

Présentation de l’information

Nous ne détaillerons pas ici tout le traitement de l’information en amont, c’est à dire avant la phase de présentation. Nombre d’ouvrages l’enseignent et, qui plus est ce n’est pas le sujet qui nous intéresse ici. Ce traitement avant présentation est effectué par l’ordinateur, et celui-ci, sauf mauvaise configuration, ne peut pas manipuler volontairement le résultat. Il exécute ce qu’on lui demande, rien de plus.

Les résultats du traitement informatisé de l’information peuvent être présentés sous deux formes différentes : les tableaux et les graphiques. Intéressons nous au contenu et à la définition de ces informations présentées de ces différentes façons.

A cette étape, les informations sont présentées sous la forme d’ « indicateurs ». Ces derniers sont les résultats de l’application d’opérations mathématiques faisant appel aux méthodes statistiques. Ces « indicateurs » servent à caractériser un groupe de données.

Ces indicateurs peuvent être regroupés en deux catégories suivant ce qu’ils expriment. Certains d’entre eux expriment une tendance centrale, d’autres une dispersion des données. Chacun de ces indicateurs offrent une vue différente mais incomplète de l’information; une tendance n’indique en rien la dispersion, et la dispersion ne représente rien si elle ne gravite pas autour d’une tendance centrale. On l’aura compris, ces indicateurs se doivent d’être présentés conjointement, mais c’est rarement le cas dans la pratique.

Revenons sur ces différentes indicateurs, un à un, découvrons leur utilité par l’illustration de plusieurs exemples, et détaillons particulièrement leurs failles.

Les indicateurs de tendance centrale

Les indicateurs de tendance centrale sont fortement utilisés, on les retrouve dans les articles de presse, les rapports d’entreprise, les débats télévisés, les sondages, etc. Comme leur nom l’indique, ils expriment une tendance, un centre, une position autour de laquelle les valeurs du groupe de données tendent à se concentrer. Ils résument l’ensemble des valeurs en une seule information qui se veut représentative. Ils facilitent ainsi l’analyse des distributions (groupes de données) au sein desquelles est présent un nombre important de valeurs.

La moyenne

La moyenne est la mesure statistique la plus courante, elle est comprise par tous. D'un point de vue mathématique, c'est le quotient de la somme des valeurs de l'ensemble des données de la distribution par l'effectif global.

$$\bar{x} = \frac{\sum x_i}{n}$$

Par exemple, le tableau suivant présente les informations du système de pointage horaire pour les employés d'une petite entreprise pour la journée du 4 Août 2015 :

Prestations journalières (04/08/2015)

Paul	4h00
John	5h00
George	7h00
Ringo	8h00

Calcul de la moyenne :

$$\bar{x} = \frac{\sum x_i}{n} = \frac{4h + 5h + 7h + 8h}{4\ employés} = \frac{24h}{4} = 6h00$$

Grâce à ce résultat, nous pouvons affirmer que les employés ont travaillé en moyenne 6h00 pour cette journée. Cela est facilement compréhensible par tout le monde car nous sommes familiarisés avec cette notion de moyenne. Pourtant, cet indicateur ne donne aucune information sur la variété – la dispersion – des valeurs étudiées. Cela va de soi après cet exemple, de dire que nous sommes incapables de savoir si tous les employés prestent leurs heures telles que prévues dans leur contrat d'embauche, si il y a des paresseux ou au contraire des super-zélés. Ce manque d'information sur la dispersion peut tronquer notre interprétation et notre avis sur le sujet étudié.

Prenons un autre exemple avec, cette fois-ci, des données réelles et vérifiées : un article de presse se basant sur les statistiques Eurostat de 2011 affirme que l'on gagne mieux sa vie au Danemark plutôt qu'aux Pays-Bas. Voici ce qui est présenté :

Salaire mensuel brut moyen en 2011

Danemark	4408 €
Pays-Bas	3451 €

Gagne-t-on mieux sa vie au Danemark qu'aux Pays-Bas? En théorie oui, si l'on

considère que tous les habitants de ces pays gagnent le même salaire. Mais en pratique ? Les Danois gagnent-ils mieux leur vie que les Néerlandais ? Probablement pas, car il est nécessaire pour affirmer cela de vérifier différents paramètres de dispersion pour ces deux nations : la répartition des classes sociales, le taux de chômage, le coût de la vie, etc. Intéressons nous par exemple, aux taux de risque de pauvreté pour ces deux pays pour la même année 2011 (source Eurostat) :

Taux de pauvreté (seuil : 60% du revenu équivalent médian)

Danemark	13,1%
Pays-Bas	10,1 %

Sur base de cette nouvelle information fournissant un soupçon de répartition des richesses, on observe que les Pays-Bas ont un taux de pauvreté significativement inférieur au Danemark. Dès lors, on voit que si les Danois gagnent en moyenne mieux leur vie que les Néerlandais, ces derniers sont moins pauvres.

L'objectif de ces exemples est de démontrer que la moyenne n'indique qu'une valeur centrale autour de laquelle les valeurs de la distribution tendent à se concentrer, et rien de plus. Même si cet indicateur est couramment utilisé et compris par tous, en représentant correctement une tendance centrale, il ne reflète pas du tout la dispersion des données et il n'est donc pas conseillé de tirer des conclusions sur base de ce seul indicateur.

Une propriété importante de cet indicateur est l'influence des valeurs extrêmes. Les valeurs extrêmes sont les valeurs considérées comme aberrantes au sein d'une distribution. Si la moyenne prend en compte toutes les données d'un ensemble pour être mesurée, elle est peu influencée par les valeurs extrêmes lorsqu'elle est mesurée au sein de grandes distributions. C'est évidemment l'inverse avec de petites distributions.

Par exemple, la moyenne des données ci-dessous est fortement influencée par une valeur extrême car la taille de la distribution est petite (l'effectif global est de six éléments). Afin d'accentuer l'influence d'une valeur extrême sur la moyenne, la valeur choisie comme 6ème élément est particulièrement aberrante :

	Élément 1	Élément 2	Élément 3	Élément 4	Élément 5	Élément 6 (valeur extrême)
Valeurs	10	15	20	25	30	1.000.000

Ici la moyenne est de 166.683,33 et il est difficile d'imaginer que tous les éléments de la

distribution tendent vers cette valeur. A l'inverse, pour une grande distribution avec un effectif d'un million d'éléments, nous observons une diminution de l'influence de cette valeur extrême sur la mesure de la moyenne :

Éléments 1 à 500.000	Éléments 500.001 à 999.999	Élément 1.000.000 (valeur extrême)
Valeur=10	Valeur=20	Valeur=1.000.000

Dans ce cas-ci, la moyenne est de 15,99 ce qui nous parait être une position centrale pertinente pour toutes les valeurs à l'exception de la valeur extrême dans notre distribution de données.

Le mode

Le mode est une mesure statistique non calculée, elle est par définition la valeur la plus fréquente au sein d'une distribution de données. Même si le terme « mode » est rarement utilisé, c'est pourtant un concept très courant qui représente la notion de majorité.

Prenons un exemple pratique où l'on demande à dix étudiants croisés en rues sur le campus universitaire de Louvain-la-Neuve, de compter le nombre de pièces d'un et de deux euros qu'ils possèdent et d'en donner le montant total. Voici les résultats obtenus :

Étudiant 1 : 3 €	Etudiant 6 : 8 €
Étudiant 2 : 5 €	Etudiant 7 : 2 €
Etudiant 3 : 2 €	Etudiant 8 : 3 €
Etudiant 4 : 1 €	Etudiant 9 : 2 €
Etudiant 5 : 5 €	Etudiant 10 : 5 €

Ici, la moyenne est de 3,60 € et le mode est égal à 2 € et à 5 €. Mesurons les fréquences des différentes valeurs :

1 €	1 étudiant
2 €	3 étudiants
3 €	2 étudiants
5 €	3 étudiants
8 €	1 étudiant

La fréquence maximale observée est de trois étudiants, cette fréquence est observée

pour deux valeurs qui sont 2 € et 5 € ; ce sont les valeurs les plus représentées au sein de cet échantillon. En effet, il est possible d'avoir plusieurs modes pour une distribution.

On l'aura compris, cette mesure statistique est très simple mais, comme la moyenne, elle ne fournit aucune information concernant la dispersion des valeurs étudiées. Mais comment peut-elle nous désorienter et fausser notre analyse?

Reprenons un exemple concret avec des donnés réelles et vérifiées ; Suite aux dernières élections de 2014 en Wallonie, le parlement wallon voit répartir ses sièges comme suit :

Répartition des sièges au parlement wallon en 2014 (75 sièges)

Gauche	30 sièges	40 %
Droite	25 sièges	33,33 %
Centre	13 sièges	17,33 %
Vert	4 sièges	5,33 %
autres	3 sièges	4 %

Le parti le plus représenté est la gauche avec 30 sièges sur 75, soit 40% de l'assemblée. C'est le mode, le parti numéro 1, celui qui a recueilli le plus grand nombre de voix, et qui détient, par conséquent, le plus grand nombre de sièges.

Imaginons à présent le résultat suivant après les futures élections (données fictives) :

Répartition des sièges au parlement wallon en 2018 (75 sièges)

Gauche	21 sièges	28 %
Droite	24 sièges	32 %
Centre	20 sièges	26.66 %
Vert	4 sièges	5,33 %
autres	6 sièges	8 %

Cette fois-ci le parti le plus représenté, le mode, est la droite avec 24 sièges. Il ne serait pas fou d'imaginer un article de presse annonçant :

« La majorité n'appartient plus à la gauche, la tendance est à la droite ! ».

Pourtant, s'il est exact d'énoncer que la tendance s'est inversée de la gauche vers la droite suite à ces nouvelles élections, cela ne veut pas dire que la droite est la grande gagnante de ces élections. Preuve en est qu'ils ont perdu un siège depuis 2014. Ils sont en tête, certes, mais moins nombreux. Pourront-ils dès lors faire tendre la politique wallonne vers leurs idées ? Ce n'est pas si sûr !

Ce qui est démontré ici, c'est que si le mode est facilement identifiable et compréhensible, et s'il est particulièrement utile pour toute analyse de type classement, il ignore néanmoins toutes les autres valeurs de la distribution et n'indique en rien si la majorité qu'il exprime est relative ou absolue (plus de 50% des valeurs). Cet exemple montre qu'il est aberrant de comparer deux distributions par leur mode.

La médiane

La médiane est la valeur située au milieu d'une distribution numérique ordonnée ; autrement dit c'est la valeur partageant les données en deux groupes de même nombre d'éléments. Elle est obtenue comme suit (n=effectif global) :

Cas d'un nombre impair d'éléments dans la distribution:

$$\tilde{x} = valeur\ de\ l'élément \frac{n+1}{2}$$

Cas d'un nombre pair d'éléments dans la distribution :

$$\tilde{x} = valeur\ moyenne\ entre\ les\ éléments\ \frac{n}{2}\ et\ \left(\frac{n}{2}\right)+1$$

Attention à ne pas confondre les notations entre la moyenne $\overline{x}$et la médiane $\tilde{x}$.

C'est également un indicateur de position, de tendance centrale. Observons son utilité sur un exemple simple : interrogeons la base de données d'une école pour connaitre les résultats des étudiants d'une classe, pour un examen précis:

Classement ascendant des étudiants en français à l'examen (sur 20 points)

Étudiant A	0
Etudiant B	3
Etudiant C	4
Etudiant D	9
Etudiant E	11
Etudiant F	13
Etudiant G	15
Etudiant H	15
Etudiant I	16
Etudiant J	17
Etudiant K	18

La moyenne des notes à cet examen est de 11/20. La valeur médiane est de 13/20. Si

la note minimale acceptable pour la réussite de l'examen et de 12/20, nous pourrions affirmer qu'en moyenne les élèves de cette classe ont échoué, puisque la moyenne n'est que de 11/20. Pourtant l'élève médian (l'étudiant F) à obtenu la note de 13/20, ce qui signifie qu'au moins la moitié des étudiants de cette classe ont réussi l'examen. En effet, sur les 11 étudiants de cette classe, 6 ont réussi.

Nous observons ici l'influence des valeurs extrêmes sur la moyenne dans le cas d'une petite distribution de 11 éléments. La moyenne est fortement diminuée par les notes des étudiants A, B, C et D qui sont particulièrement mauvaises même si il y a beaucoup de bonnes notes.

Par contre, l'avantage de la médiane est de ne pas tenir compte de ces valeurs extrêmes et se révèle dans ce cas-ci beaucoup plus utile, même si elle ne tient pas compte de la dispersion des données au sein de la distribution. En effet, elle indique bien qu'une part plus importante des étudiants ont réussi. Soulignons également que cette mesure est beaucoup moins utilisée que la moyenne et n'est probablement pas aussi facilement comprise par tout le monde. Et pourtant dans le cas présent elle serait plus utile pour caractériser notre distribution.

Reprenons l'exemple concret du salaire au Danemark et aux Pays-Bas. Selon Eurostat, le revenu annuel médian pour ces 2 pays en 2011 est :

Revenu annuel médian 2011

Danemark	26.421 €
Pays-Bas	20.310 €

Une fois de plus le Danemark l'emporte haut la main. Cependant nous n'avons toujours pas d'idée sur la répartition des richesses au sein de ces deux populations. Nous connaissons cependant la différence du taux de pauvreté qui nous indique que le taux de Néerlandais sous le seuil de pauvreté est plus faible qu'au Danemark.

A l'inverse, selon le classement annuel des têtes les plus riches du monde en 2011 par le magazine Forbes, la tête la plus riche des Pays-Bas pèse 7,5 millions de dollars, alors que celle du Danemark ne pèse que 4,7 millions de dollars. Autrement dit la valeur extrême supérieure est beaucoup plus forte aux Pays-Bas qu'au Danemark.

Une fois de plus, ce n'est pas parce que la valeur médiane est faible que la répartition au sein d'une distribution l'est également. Avec ce que nous avons pu obtenir à présent comme informations, nous savons qu'il y a moins de pauvres aux Pays-Bas qu'au Danemark et que la personne la plus riche de ces 2 pays est Néerlandaise, ce qui peut paraître contradictoire par rapport à ce que le revenu mensuel moyen et le revenu annuel médian peuvent laisser penser.

Les indicateurs de dispersion

Les indicateurs de dispersion sont très utiles, ils sont pourtant très peu utilisés en BI et malheureusement ne sont pas toujours compris par tous les intervenants. Ils représentent la variété, la dispersion des différentes valeurs au sein d'une distribution.

L'écart moyen

L'écart moyen est la moyenne des écarts par rapport à la valeur moyenne de la distribution. Cette mesure statistique ne représente rien à elle seule, elle doit toujours être accompagnée de la moyenne. C'est sans doute l'indicateur qui représente le mieux la dispersion des résultats puisqu'il est calculé à partir de l'écart entre chaque valeur de la distribution et la moyenne.

$$e_m = \frac{\sum |\overline{x} - x_i|}{n}$$

Prenons par exemple le résultat de la finale féminine du 100 mètres lors des derniers jeux olympiques de Londres en 2012 :

Résultats finale du 100m féminin, JO 2012 de Londres

Shelly-Ann Fraser	10s75
Carmelita Jeter	10s78
Veronica Campbell-Brown	10s81
Tianna Madison	10s85
Allyson Felix	10s89
Kelly-Ann Baptiste	10s94
Murielle Ahouré	11s00
Blessing Okagbare	11s01

La moyenne de cette distribution est facilement calculable. Un 100m féminin est effectué en moyenne en 10s88 par ces athlètes. Calculons à présent les écarts absolus des résultats pour chacune d'elles :

Participante	*Chronomètre*	$\lvert\overline{x} - x_i\rvert$
Shelly-Ann Fraser	10s75	|10s88-10s75|= 0s13
Carmelita Jeter	10s78	|10s88-10s78|= 0s10
Veronica Campbell-Brown	10s81	|10s88-10s81|=0s07
Tianna Madison	10s85	… 0s03
Allyson Felix	10s89	… 0s01

Kelly-Ann Baptiste	10s94	… 0s06
Murielle Ahouré	11s00	… 0s12
Blessing Okagbare	11s01	… 0s13

Appliquons à présent la formule de l'écart moyen :

$$e_m = \frac{\sum |\overline{x} - x_i|}{n} = \frac{0s13 + 0s10 + 0s07 + 0s03 + 0s01 + 0s06 + 0s12 + 0s13}{8} = 0s08$$

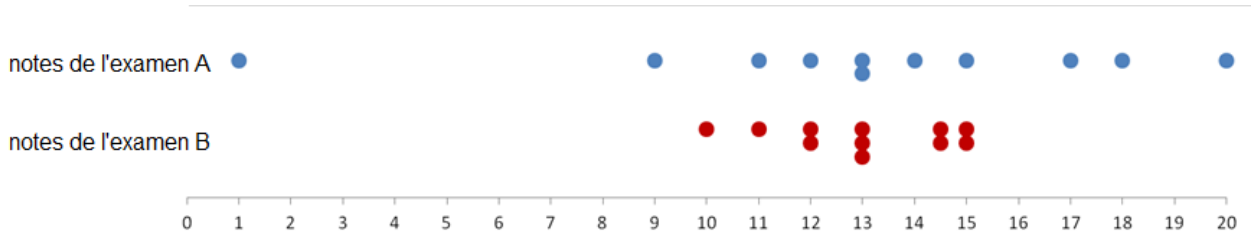
Ceci nous indique que le 100m féminin est effectué en moyenne avec un écart de 0s08 par rapport à la moyenne globale de 10s88. Autrement dit les résultats s'écartent en moyenne de 0s08 par rapport à la moyenne 10s88. Si nous reprenons les résultats et vérifions combien de participantes sont comprises dans l'intervalle $\overline{x} - e_m$ et $\overline{x} + e_m$, soit entre 10s80 et 10s96, nous comptons 4 participantes sur les 8, soit 50% des participantes.

Il est évident que cet indicateur de dispersion ne représente rien à lui seul. En reprenant notre classe d'étudiants et leurs notes pour deux autres examens, observons la dispersion des valeurs à l'aide de l'écart moyen:

Examen A	Examen B
1/20	10/20
9/20	11/20
11/20	12/20
12/20	12/20
13/20	13/20
13/20	13/20
14/20	13/20
15/20	14,5/20
17/20	14,5/20
18/20	15/20
20/20	15/20

Notons tout d'abord que les valeurs moyennes, modes et médianes de ces deux examens sont égales à 13/20.

L'écart moyen pour les notes obtenues à l'examen A est de 3,5/20 alors que l'écart moyen des notes obtenues à l'examen B est de 1,3/20. Les notes de l'examen B sont donc nettement moins dispersée que celles de l'examen A, nous pouvons l'observer facilement via ce graphique reprenant toutes les notes obtenues aux deux examens:



Il est important de rappeler que l'écart moyen ne représente rien s'il n'est pas accompagné par la moyenne. En effet, si pour un troisième examen l'écart moyen des notes obtenues est de 0/20, cela signifie qu'il n'y a aucune dispersion parmi les notes, sans spécifier si tout les étudiants ont échoués avec 4/20 ou, au contraire, réussi avec 18/20.

Si l'écart moyen donne une bonne représentation de la dispersion des données au sein d'une distribution, il n'offre aucune information sur la répartition de cette dispersion. Par exemple, on peut observer via le tableau et le graphe précédent que les notes supérieures (au dessus de la moyenne) obtenues à l'examen A sont fortement concentrées contrairement aux notes inférieures qui sont beaucoup plus éparses. Pour compléter notre analyse en prenant compte de cette répartition de la dispersion, intéressons-nous aux quartiles et à l'intervalle interquartiles.

Mais avant cela, notons également qu'il existe une autre mesure, l'écart type qui, tout comme l'écart moyen, mesure la dispersion des données autour de la moyenne. Mais contrairement à l'écart moyen qui mesure la moyenne des écarts par rapport à la moyenne de la distribution, l'écart type mesure la moyenne quadratique des écarts par rapport à la moyenne de la distribution (la moyenne des carrés des écarts).

$$s = \sqrt{\frac{\sum(\overline{x} - x_i)^2}{n-1}}$$

L'écart type est la mesure la plus utilisée en statistique et pourtant elle est la moins comprise par les non initiés. Elle sert notamment pour calculer les corrélations entre deux distributions, autrement dit le lien de cause à effet entre deux phénomènes étudiés. Elle sert aussi, dans les sondages d'opinions, à évaluer l'incertitude des variations accidentelles, ce qui est appelé « marge d'erreur » dans ce cas. Nous ne détaillerons pas ici cette mesure car elle est rarement utilisée dans un processus BI.

L’intervalle interquartile

L’intervalle interquartile d’une distribution est, par définition, l’intervalle qui comprend les 50%, soit la moitié, des données les plus au centre de cette distribution. Il est égal à la différence entre deux mesures statistiques que sont les quartiles.

$$IQ = Q_3 - Q_1$$

Le premier quartile (Q_1) est la valeur pour laquelle 25% des valeurs de la distribution lui sont inférieures, et par conséquent, 75% des valeurs lui sont supérieures.

Le troisième quartile (Q_3) est la valeur pour laquelle 75% des valeurs de la distribution lui sont inférieures, et par conséquent, 25% des valeurs lui sont supérieures.

Il existe un quartile Q_2 qui divise la distribution en deux groupes égaux en nombre (50%), ce deuxième quartile est la médiane.

Ces quartiles sont particulièrement utiles lorsqu’ils sont représentés avec la médiane, le minimum et le maximum dans une représentation graphique appelée box-plot. C’est un diagramme en boite représentant l’intervalle interquartile, entouré de deux segments de droite dont les extrémités sont les valeurs minimum et maximum de la distribution. Ceci offre une représentation rapide et résumée de la dispersion des valeurs où les 4 quarts sont facilement identifiables.

Prenons comme exemple une entreprise e-commerce de vente de produits de soins et de santé. Elle souhaite apprécier le taux de satisfaction de ses clients en se basant sur les résultats d’une enquête, résultats pré calculés et exprimés en pourcentage de satisfaction. Voici les résultats de l’enquête et leur représentation box-plot:

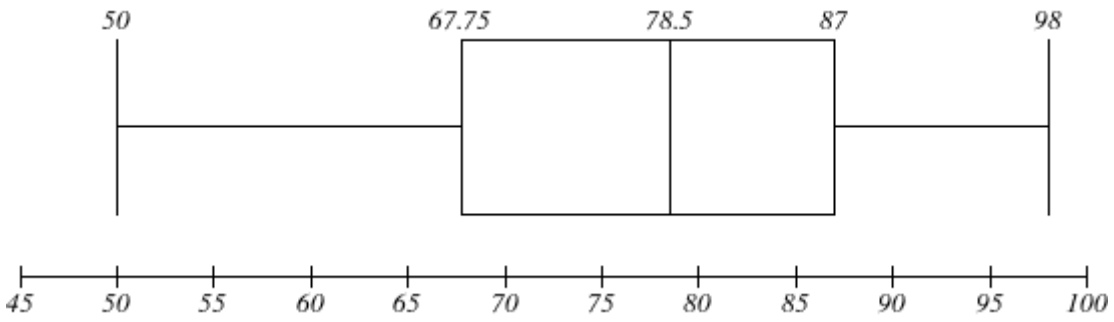
Note minimale obtenue = 50%

Q_1= 67,75 %

$Q_2 = \tilde{x}$ = 78,5 %

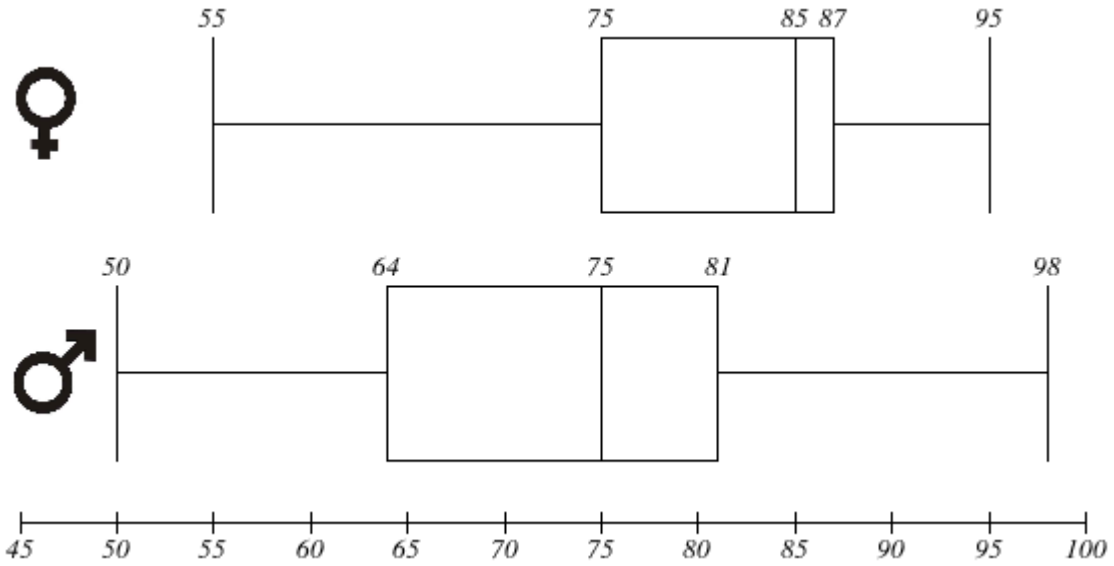
Q_3= 87 %

Note maximale obtenue = 98%



Nous observons ici une dispersion relativement homogène des données, même si les 25% les moins satisfaits des clients représentés par le premier segment (entre 50 et 67,75% de satisfaction) ont un avis légèrement plus dispersé que les 25% de clients les plus satisfaits, représentés par le dernier segment; le premier segment de droite étant plus long que le dernier. On peut également observer le même phénomène en comparant les deux boites au centre du graphique, cependant la différence n'étant pas spectaculaire, nous considérerons cette dispersion homogène. Si l'objectif de cette entreprise est un taux de satisfaction client supérieur à 75%, c'est une réussite car la valeur médiane est de 78,5%. Plus de la moitié des clients sont en effet satisfaits.

Qu'en est-il si l'on distingue à présent les avis de la clientèle féminine des avis de la clientèle masculine, et ce dans des graphes semblables ?



On observe une nette différence dans la dispersion des deux distributions ; Si les 3/4 des clientes sont satisfaites à plus de 75%, seule la moitié des clients masculins le sont. On observe également que la répartition des valeurs dispersées pour les clients masculins est relativement homogène avec, cependant, une concentration entre la médiane et le troisième quartile. En ce qui concerne les clientes, il y a une forte concentration entre la médiane et le troisième quartile alors que les avis les plus négatifs sont beaucoup plus éparses ce qui en fait une distribution dont la dispersion est hétérogène.

Avec une représentation box-plot, nous obtenons donc une vue d'ensemble rapide de la dispersion des données pour une ou plusieurs distributions. Nous pouvons d'ailleurs facilement comparer plusieurs distributions entre elles. Si une représentation box-plot est fort utile, elle nécessite cependant une connaissance de la représentation pour être comprise.

Avant d'aller plus loin

Dans cette première partie, on a pu appréhender quelques indicateurs statistiques qui permettent de compléter notre analyse de l'information. Il ne s'agit cependant que d'une introduction car il existe bien d'autres indicateurs, nous n'avons parcouru que les plus couramment utilisés.

La seconde partie, qui concerne l'interprétation de l'information, se concentre moins sur les aspects mathématiques tels que nous les avons vus jusqu'à présent. Nous parlerons plutôt de plusieurs caractéristiques de l'information pouvant influencer le processus mental de l'interprétation, sans pour autant basculer dans une étude psychologique mais en nous focalisant sur un raisonnement rationnel.

Interprétation de l’information

La plus grande faille, le plus grand risque d’erreur, de manipulation du processus décisionnel réside dans la phase d’interprétation de l’information. Dans un premier temps, l’information est présentée sous sa forme objective ; il s’agit d’un résultat, d’une conclusion mathématique obtenue après traitement. Il est important ici d’obtenir une information correcte faisant suite à plusieurs calculs suivants des formules démontrées comme c’est le cas en statistique. L’homme de science exacte ne peut se permettre de fausser son résultat, que ce soit par modification de formules ou modification des données elles-mêmes. Le traitement de l’information ne peut inclure que des processus mathématiques avérés. Ils conduisent aux calculs des indicateurs analysés dans le chapitre précédent.

Lorsque l’information est présentée, énoncée et entendue par l’homme, le cerveau humain met en route un processus de raisonnement et l’information n’est plus simplement traitée mais interprétée. L’homme est ainsi fait, il ne stocke pas l’information de manière brute comme un ordinateur mais il fait plutôt appel à ce qu’il connait pour rattacher cette information à une autre déjà assimilée. Elle est amalgamée aux sentiments, aux valeurs, à l’inconscient de la personne pour en ressortir interprétée, reformatée.

Sauf si cet esprit est cartésien, hautement pragmatique, et s’il connait en détails les dangers de la statistique, l’interprétation qu’il effectuera se retrouvera incomplète et dans le pire des cas, faussée. Dans un processus décisionnel, ce phénomène peut être dévastateur.

La décision est l’aboutissement du raisonnement qui, à partir d’une information présentée, détermine une information interprétée en appliquant des lois de transformation. Ces lois de transformations sont de deux types ; elles peuvent être de type mathématique (faisant suite par exemple à la compréhension des indicateurs du chapitre précédent) ou de type empirique. C’est précisément ces lois de transformation de type empirique qui sont le plus souvent à l’origine d’erreurs et de mauvaise interprétation de l’information, par manque d’expérience et par manque de connaissance approfondie du sujet étudié. Nous allons étudier maintenant deux concepts de transformation de l’information qui, lorsqu’ils ne sont pas correctement appliqués, peuvent fausser le raisonnement humain, l’information interprétée et la décision qui en découle.

Le poids relatif de l'information

Toute information numérique possède un poids relatif. Ce poids correspond à l'importance de cette information vis-à-vis d'une information de référence. Par exemple, le salaire net d'un employé à un poids relatif plus ou moins important vis-à-vis de son salaire brut, le coût d'un bien à un poids relatif vis-à-vis du budget qu'il lui est alloué, etc. Ce qui est important c'est de concevoir cette notion de poids relatif et d'admettre qu'il peut varier, par exemple dans le temps, sans pour autant changer la valeur de l'information initiale.

Imaginons la somme de 10.000 euros placée en bourse. La première semaine, les actions achetées grimpent de 20%. La deuxième semaine, elles grimpent à nouveau de 20%, mais malheureusement la troisième semaine elles accusent une perte de 40%. Qu'en est-il au final après ces trois semaines ?

10.000€ + 20% + 20% - 40% = ?

Penser que la somme soit revenue à son état initial est une erreur. Reprenons pas à pas la valeur de ce portefeuille d'actions :

1er jour : *10.000 €*

Après une semaine, hausse de 20% :

*10.000€ + (20% * 10.000€) = 10.000€ + 2.000€ = 12.000€*

Une semaine plus tard, à nouveau une hausse de 20% :

*12.000 € + (20% * 12.000€) = 12.000€ + 2.400€ = 14.400 €*

La dernière semaine, chute de 40% :

*14.400€ - (40% * 14.400 €) = 14.000€ - 5.760€ = 8640 €*

Soit une perte de plus 1360 euros par rapport à l'investissement de départ !

Même si l'on peut de prime abord considérer que la chute de 40% ne vient qu'équilibrer les deux hausses de 20%, il est utile de considérer la dimension temporelle qui altère chaque semaine le poids de notre information. Deux modifications similaires ont un impact différent suivant le poids relatif de l'information à laquelle elles s'appliquent. Dans ce cas-ci, le poids relatif de l'information est différent chaque semaine. En effet, si la première hausse de 20% est relative à l'information initiale 10.000€, ce n'est plus le

cas de la deuxième hausse qui est relative à 10.000€+20% (et non plus 10.000€). Enfin, la dernière altération, la chute de 40%, est relative à 12.000€+20% (et non pas à 10.000€, ni même à 10.000€+40%).

Le poids de l'information est une caractéristique très importante qui permet d'ajouter une valeur relative à cette information. Illustrons cela avec un autre exemple, le journal Le Soir publia cet article le jeudi 10 avril 2008 :

> *Bigot quitte la RTBF pour Endemol France*
>
> *Le Français Yves Bigot, directeur des antennes de la RTBF depuis avril 2006, quittera la télévision publique le 1er septembre prochain. Il a en effet trouvé un accord avec Endemol France dont il prendra la direction des programmes.*
>
> *… … …*
>
> *Quant au bilan de Bigot à Reyers, il reste globalement très positif malgré quelques échecs. C'est en tout cas l'avis de Jean-Paul Philippot : « La Une a stabilisé ses audiences, la Deux les a augmentées de 43%. On a investi dans la production belge, comme Melting-pot café, lancé Arte Belgique, des talk-shows, le 12 minutes et mis Matin première en télé. Surtout, la RTBF a renoué avec des événements et avec une politique remarquable en termes d'acquisition de fictions et de droits sportifs. Nous aurions pu travailler ensemble longtemps encore. »*

L'information objective énoncée ici est une mesure de médiamétrie, d'audimat d'audiences. Il est dit que la chaine de télévision « la Deux » a augmenté ses audiences de 43%. C'est un très bon résultat ! Mais qu'en est-il réellement ?

Selon le site web du conseil supérieur de l'audiovisuel (CSA), la Deux enregistrait 3% d'audimat fin 2005 et 4,5% fin 2007. Ceci correspond à une progression de 50%, les 43% énoncés correspondent à la progression entre les mesures non arrondies. Néanmoins, cela reste une part de marché ridicule face aux concurrents qui affichaient pour le meilleur d'entre eux une part de marché de 20% fin 2007.

Dans ce cas-ci, seul le poids relatif de la progression entre deux informations objectives est énoncé. Les données objectives ne sont pas présentes et il est impossible de juger de la qualité de cette progression. Il manque également une notion de paysage étudié ; Dans quel environnement vivent ces informations ? Que représentent-elles dans ce paysage ? Quelles sont les autres éléments qui constituent cet environnement ?

Ce sont quelques questions pour lesquelles on se doit d'y répondre pour que notre

phase d'interprétation passe avant tout par une bonne analyse. En obtenant les mesures médiamétriques pour 2005 et 2007 pour cette chaine de télévision mais également pour les autres chaines concurrentes, le paysage étudié se complète et notre interprétation change par l'observation très « relative » de cette progression de 43% ; Il est beaucoup plus facile pour l'audimat de « La Deux » de progresser à ce point avec un audimat de départ aussi faible, que pour une chaine concurrente partant avec un audimat de départ de 20%. Qui plus est, l'information est ici énoncée avec enthousiasme par l'interviewé qui communique sans détour son point de vue très positif. C'est un jeu dangereux où le processus d'interprétation par le lecteur est bridée. Ce dernier peut être bloqué dans son processus d'interprétation propre par l'influence de l'expression très positive des résultats dans l'article.

Le poids relatif de l'information est une notion essentielle. Nous avons vu l'impact qu'il occasionne si il change au cours du temps, et la nécessité de le prendre en compte pour la comparaison d'une information avec d'autres.

La déduction par corrélation

La corrélation est l'expression de relation par parallélisme, par comparaison. C'est le lien de dépendance d'une information envers une ou plusieurs autres informations. Ce lien est d'influence, de cause à effet, et cela a parfois un impact fort ou au contraire limité sur l'information première.

Prenons un exemple simple où le taux de satisfaction des clients d'une piscine en plein air serait fortement dépendant des conditions climatiques. Il est facile d'imaginer qu'un ciel bleu enchanterait nos clients, par contre un temps gris et pluvieux impacterait fortement leurs avis. Si la qualité de service, le prix et tout autre paramètre est resté stable, nous pouvons supposer l'existence d'une corrélation relativement élevée entre la satisfaction des clients et la météo.

Si en statistique la relation entre deux informations peut se mesurer à l'aide du coefficient de corrélation, au niveau de l'interprétation c'est une notion différente qui se doit d'être déterminée par déduction, par réflexion, en jugeant de la pertinence d'une telle relation entre les informations. Nous allons voir par un exemple concret, en quoi la corrélation est une notion importante pouvant fausser l'interprétation de l'information.

L'agence de presse Belga News publiait en octobre 2013 l'article suivant :

> *Consommation de Champagne : Le Japon dépasse la Belgique*
>
> *Environ 8,3 millions de bouteilles de vins de Champagne ont été exportées vers la Belgique en 2012, selon les chiffres présentés jeudi lors d'une conférence de*

presse par le Bureau du Champagne du Benelux. Si les expéditions vers la Belgique apparaissent en baisse par rapport aux années 2011 (9,6 millions de bouteilles) et 2010 (8,8 millions), le pays reste un consommateur historique et connaisseur des vins effervescents en provenance de Champagne, selon Grégoire Van den Ostende, directeur du Bureau du Champagne Benelux.

"*Les chiffres d'exportation officielle sont en légère baisse, mais nous savons par ailleurs que de plus en plus de Belges se rendent sur place pour acheter des bouteilles chez le producteur, et ces achats-là ne sont pas comptabilisés dans les expéditions*", *précise Grégoire Van den Ostende. Ces achats sur place seraient au nombre de 2 à 5 millions de bouteilles, plus probablement aux alentours de 3-4 millions, selon les propos de Grégoire Van den Ostende.*

La Belgique perd une place dans le classement

*Les vins de Champagne, dont les cépages proviennent des départements de la Marne, l'Aube, la Haute-Marne, l'Aisne et la Seine-et-Marne, sont exportés dans plus de 190 pays. Si la Belgique figurait ces dernières années à la 4e place parmi les plus importants marchés extérieurs importateurs de "*bulles*", elle recule en 2012 d'une place, dépassée par le Japon qui s'installe derrière le trio de tête habituel.*

Les trois premières places restent inchangées, avec le Royaume-Uni (32,4 millions de bouteilles), les Etats-Unis (17,7 millions) et l'Allemagne (12,6 millions). La Belgique et ses 8,3 millions de bouteilles importées représente malgré tout à elle seule 6% du total des exportations.

Cet article énonce dans son titre que le Japon consomme plus de Champagne que la Belgique. Cependant, le contenu même de l'article n'est que très peu pertinent vis-à-vis de cette idée. Qu'en est-il vraiment ? Jetons tout d'abord un œil sur les chiffres présentés dans le contenu de l'article. Ces chiffres concernent l'exportation de Champagne en provenance de France vers plusieurs pays pour l'année 2012. Un classement des marchés extérieurs importateurs est réalisé sur base de ces chiffres :

1er. Royaume-Uni avec 32,4 millions de bouteilles

2ème Etats-Unis avec 17,7 millions de bouteilles

3ème Allemagne avec 12,6 millions de bouteilles

4ème Japon mais le chiffre exact inconnu

5ème Belgique avec 8,3 millions de bouteilles

Nous ne connaissons pas ici le montant de bouteilles importées par le Japon mais une recherche rapide sur Internet nous renseigne les chiffres officiels du Comité Interprofessionnel du Vin de Champagne (CIVC) et présente la somme de 9.062.972 bouteilles exportées vers le Japon pour 2012. Notons que ce comité présente les mêmes résultats que dans l'article pour les autres pays.

Le Japon est-il dès lors 4ème au classement ? Oui, au classement des marchés extérieurs importateurs de Champagne. Mais le titre de l'article ne parle pas d'importation mais de consommation de bouteille de Champagnes. Le Japon consomme t-il réellement plus de Champagne que la Belgique ?

Le contenu de l'article ci-dessus de même que le CIVC l'affirme, entre 3 et 4 millions de bouteilles de Champagne sont achetées directement en France par les Belges. Dès lors, importation égale à consommation ? Non, c'est une corrélation incomplète et fausse dans ce cas-ci. Les données présentées concernent les exportations de Champagne, sur base desquelles il est correct d'effectuer un classement par marchés, par pays. Dans ce cas, il est tout à fait correct de ne pas prendre en compte l'achat direct en France.

Mais si l'on parle de consommation, ne faudrait-il pas considérer l'achat direct en France ? Si il est facilement concevable d'affirmer que seul le champagne importé au Japon est consommé au Japon, ce n'est pas la même conclusion pour la Belgique où, en plus des 8,3 millions importés, se rajoute entre 3 et 4 millions de bouteilles en achat direct. Nous pouvons donc raisonnablement conclure (sans statistique vérifiée cependant) que les Belges consomment plus de champagne que les Japonais. Il est probable que les Allemands et les Anglais, faisant également partie de la même zone de libre échange que les Belges, achètent également une quantité non négligeable de bouteilles de Champagne en achat direct chez les producteurs Français.

La corrélation exprimée par le titre de l'article entre importation et consommation est directe, or il est démontré au sein même de cet article que l'achat direct n'est pas négligeable. La relation entre importation et consommation n'est donc que partielle. Si tout ce qui est importé est consommé, il est faux de considérer l'importation comme le seul facteur participatif à la consommation.

La corrélation entre deux informations est, tout comme le poids, une caractéristique ajoutant une notion de relativité. Si la plupart du temps cette relativité n'est pas prise en compte dans notre raisonnement et la relation est considérée comme directe, cela fausse notre interprétation de l'information.

Est-il correct de comparer le niveau de vie des habitants de plusieurs pays en se basant uniquement sur le revenu moyen ou médian annuel ? Peut-on juger la qualité de service d'un établissement hôtelier sur base seule du taux de satisfaction de leur

clientèle ? Le lien de cause à effet entre ces deux informations est-il direct ou relatif ? Quelle proportion à cette relativité ? Quels peuvent être les autres facteurs impactants notre résultat, quels impacts ont-ils?

Nous avons observé ici l’erreur obtenue en ne tenant pas compte de l’impact relatif d’une information sur une autre.

Conclusion

Si la première partie de ce document reprend différents concepts mathématiques pouvant présenter des failles dans l'analyse de l'information, la deuxième partie comporte d'autres notions qui ne font pas nécessairement partie du raisonnement mathématique mais du raisonnement empirique pour aboutir à l'interprétation de l'information.

L'objectif n'est bien entendu pas de profiter des différentes failles qui ont été détaillées, ni de manipuler notre interlocuteur en omettant toutes notions de relativité pour l'information. L'objectif est de connaitre ces différents concepts, de les assimiler et en avoir pleine conscience lorsque nous sommes face à de l'information dans un processus BI décisionnel.

Sources et références

La statistique en clair. Editions ellipses.
2011

Eléments de statistique. Editions de l'université libre de Bruxelles.
2008

Top 20 Dans quel pays d'Europe gagneriez-vous le plus ?. www.express.be.
2011

Taux de risque de pauvreté par seuil de pauvreté, âge et sexe. Eurostat.
2015

Elections 2014 : Parlement wallon, répartition des sièges. Elections2014.belgium.be.
2014

Revenu moyen et médian par type de ménage. Eurostat.
2015

European Billionaires Of 2011. Forbes.com.
2011

Athlétisme aux Jeux Olympiques de 2012 – 100m femmes. Fr.wikipedia.org.
2015

Affective computing and sentiment analysis. Springer Netherlands.
2011

Bigot quitte la RTBF pour Endemol France. Le Soir.
2008

Télévision : Audience et parts de marché. Csa.be.
2012

Consommation de Champagne : le Japon dépasse la Belgique. Rtbf.be.
2013

*Les expéditions de vins de champagne en 2012.*Comité Champagne CIVC.
2013

Présentation de l’auteur

Quentin Braham est un analyste senior en Business Intelligence faisant preuve d’une connaissance approfondie en plateformes de reporting. Du haut de son expérience, il est désireux d’apporter la meilleure solution tout en préservant la consistance entre les processus business et la plus-value de transformer ces derniers en informations précises et utiles.

Sa passion pour les statistiques, l’utilisation et l’impact qu’on ces dernières sur notre société, lui donne quotidiennement l’envie de partager ses connaissances.