Chapitre 4
Les tests d'hypothèses

1. Vocabulaire lié aux tests d'hypothèses

En statistiques, les réponses ne sont jamais tranchées. Il y a toujours une part d'aléatoire, d'imprévu, d'écart. C'est normal car sans cette part, les statistiques ne seraient pas nécessaires. Si un événement est certain, il n'y a pas besoin de loi de probabilités ou de test pour le prédire.

En statistiques, les tests, ou plus précisément **les tests d'hypothèses**, sont un outil, une règle de décision, qui permet de rejeter ou non une hypothèse, une réponse avec un risque connu. Le risque zéro dans l'absolu n'existe que dans les domaines où les statistiques ne sont pas nécessaires.

Le terme "hypothèse" est important. Tous les tests réalisés partent d'une affirmation à confirmer ou à infirmer. Par exemple, avec le test de Shapiro-Wilk, l'affirmation est : les données suivent une loi normale. Cette affirmation est appelée **hypothèse nulle** et écrite H_0. Rejeter H_0 revient à accepter l'**hypothèse alternative** H_1 : les données ne suivent pas une loi normale.

Le test d'hypothèses permet de dire "À 95 % (ou 90 % ou 99 %), la réponse est..." Cela signifie qu'il y a une possibilité de se tromper. Deux types d'erreurs existent : l'**erreur de première espèce** et l'**erreur de deuxième espèce** (cf. figure 04-01). L'erreur de première espèce survient lorsque H_0 est rejetée alors qu'elle est vraie. La probabilité associée à cette erreur est le **seuil de significativité**, α, généralement fixé à 5 %, soit $\alpha = 0{,}05$. L'erreur de deuxième espèce survient lorsque H_0 n'est pas rejetée ou est gardée alors qu'elle est fausse. La probabilité associée à cette erreur est liée à la **puissance du test**, $1 - \beta$ qui calcule la probabilité de rejeter H_0 alors qu'elle est bien fausse avec β l'erreur de deuxième espèce.

Remarque

Le choix entre l'hypothèse nulle et alternative ne dépend que de la question posée et du contexte, même s'il est souvent conseillé de minimiser en priorité l'erreur de première espèce. Le but est souvent de rejeter H_0 de manière certaine.

Test d'hypothèses		**Réalité**	
		H_0 est vraie	H_0 est fausse
Décision	H_0 est conservée	**Vrais positifs** **Décision juste** *H_0 est acceptée alors qu'elle est vraie* Seuil de confiance : Probabilité = $1 - \alpha$	**Faux positifs** **Erreur de 2e espèce** *H_0 est acceptée alors qu'elle est fausse* Probabilité = β
	H_0 est rejetée	**Faux négatifs** **Erreur de 1re espèce** *H_0 est rejetée alors qu'elle est vraie* Seuil de significativité : Probabilité = α	**Vrais négatifs** **Décision juste** *H_0 est rejetée alors qu'elle est fausse* Puissance du test : Probabilité $1 - \beta$

Le seuil de significativité α – ou le seuil de confiance $1 - \alpha$ – est fixé en amont par la personne réalisant le test, mais il peut aussi être fixé pour répondre à des normes industrielles, pharmaceutiques…

La puissance du test $1 - \beta$ est calculée après la réalisation du test à partir de données (moyenne, dispersion et taille des échantillons), du seuil de significativité α et du type de test d'hypothèses réalisé :

– La puissance du test est corrélée à la population étudiée, c'est-à-dire que si les groupes testés sont très différents, la puissance du test est meilleure.

Par exemple, le test d'hypothèses cherche à voir si la taille moyenne des femmes est identique à la taille moyenne des hommes. La puissance du test est dépendante de la moyenne et de la variabilité des tailles mesurées au sein de chaque genre. Si les deux genres ont des moyennes très différentes comme c'est le cas en Suède (femmes = 167 cm et hommes = 182 cm), la puissance du test est meilleure qu'à Singapour où la différence entre les genres est plus faible (femmes = 160 cm et hommes = 171 cm).

– La puissance du test est positivement corrélée à la taille de l'échantillon, c'est-à-dire que plus la taille de l'échantillon analysé est grande, plus la puissance du test augmente pour une même population étudiée.
– La puissance du test est aussi positivement corrélée à la valeur du seuil de significativité, c'est-à-dire que plus il y a d'erreurs de première espèce, moins il y a d'erreurs de deuxième espèce pour une même population.

Pour augmenter à la fois le seuil de confiance $1 - \alpha$ (donc diminuer le seuil de significativité α) et la puissance du test $1 - \beta$, il faut augmenter la taille de l'échantillon, mais la population en elle-même aura toujours une erreur liée à ses données, et le test d'hypothèses une erreur liée à sa nature. De plus, l'augmentation de la taille de l'échantillon est parfois limitée par des caractéristiques physiques (nombre d'individus disponibles), techniques (temps de la reproduction) ou financières (coût de séquençage d'ADN).

Souvent, les études médicales et biologiques donnent la **sensibilité** et la **spécificité** d'un test. La sensibilité est la capacité à mesurer la proportion de vrais positifs si H_0 est vraie, alors que la spécificité mesure la proportion de vrais négatifs si H_0 est rejetée :

- $Sensibilité = \frac{Vrais\ positifs}{Vrais\ positifs + Faux\ négatifs} = 1 - \alpha$
- $Spécificité = \frac{Vrais\ négatifs}{Vrais\ négatifs + Faux\ positifs} = 1 - \beta$

Cas d'utilisation de la puissance statistique sur les autotests COVID-19

Pour qu'un autotest COVID-19 soit autorisé, la HAS (Haute Autorité de Santé) impose qu'il soit en mesure de détecter la présence du virus chez une personne effectivement malade dans 80 % des cas a minima, et qu'ils soient en mesure de constater l'absence de virus chez une personne effectivement non malade dans plus de 99 % des cas.

Un nouvel autotest sort sur le marché et communique les résultats suivants :

	COVID détecté par PCR	**COVID non détecté par PCR**
COVID détecté par autotest	1236	35
COVID non détecté par autotest	6	1649

L'autotest détecte la présence de virus dans $\frac{1236}{1236+6} = \frac{1236}{1242} = 0,995$ = 99,5 % des cas, donc dans plus de 99 % des cas confirmés par PCR.

L'autotest détecte l'absence de virus dans $\frac{1649}{1649+35} = \frac{1649}{1684} = 0,979$ = 97,9 % des cas, donc dans plus de 97 % des cas infirmés par PCR.

Bien que ce test semble très efficace, il ne respecte pas les conditions définies par la HAS et ne pourra pas être mis sur le marché. En effet, le seuil de détection d'absence de virus est inférieur à l'attendu de 99 %.

Les tests d'hypothèses peuvent être utilisés pour valider une technique ou un outil (ici, un autotest COVID), le respect de norme de production, mais aussi la comparaison entre groupes de données. Plus d'exemples sont fournis dans les chapitres La comparaison à une valeur théorique et La comparaison de deux groupes.

Le seuil de significativité et la puissance du test peuvent être fixés en amont grâce à des données fictives pour connaître le nombre minimal d'individus statistiques nécessaires à la réalisation de l'étude. Dans ce cas, la puissance statistique minimale est généralement supérieure à 80 %, voire à 90 %.

Calcul de la puissance de deux tests

Deux tests A et B sont envisagés pour analyser les prélèvements. Le risque de première espèce est fixé à $\alpha = 0{,}05$ et le risque de deuxième espèce est calculé : 0,21 pour A et 0,14 pour B. Quel est le meilleur test ?

Il faut commencer par calculer la puissance de chacun des tests : $P_A = 1 - \beta_A = 1 - 0{,}20 = 0{,}80$ et $P_B = 1 - \beta_B = 1 - 0{,}14 = 0{,}86$. Pour une même erreur de première espèce α, la puissance est plus grande (ou l'erreur de deuxième espèce β est plus petite) pour le test B, donc il est meilleur.

Remarque

Le seuil de significativité revient à rejeter une partie des données vraies mais peu courantes (cf. figure 04-01).

2. Démarche du test d'hypothèses

La démarche du test d'hypothèses est toujours la même, quel que soit le test d'hypothèses réalisé :

– 1re étape : Définir les hypothèses.

Cas de l'autotest COVID : l'autotest permet de déterminer la présence de COVID ou l'autotest permet de déterminer l'absence de COVID. Afin de minimiser l'erreur de première espèce en suivant les attendus de la Haute Autorité de santé, l'autotest permet de déterminer l'absence de COVID.

– 2e étape : Formaliser les hypothèses.

Cas de l'autotest COVID :

H_0 = La COVID n'est pas présente dans l'échantillon.

H_1 = La COVID est présente dans l'échantillon.

– 3e étape : Définir le seuil de significativité (α) et la puissance minimale acceptée ($1-\beta$). Il faut aussi définir le type de test : bilatéral, unilatéral à droite ou à gauche. Le type de test est aussi dépendant de la statistique de test choisie.

Cas de l'autotest COVID :

$1-\alpha = 0{,}99$, donc $\alpha = 0{,}001$.

$1-\beta = 0{,}080$, la puissance minimale acceptée.

– 4e étape : Calculer la **statistique de test** et la probabilité que H_0 soit vraie en fonction des données et des hypothèses définies pour savoir si l'hypothèse nulle est rejetée et/ou calculer la sensibilité et la spécificité pour valider le test réalisé.

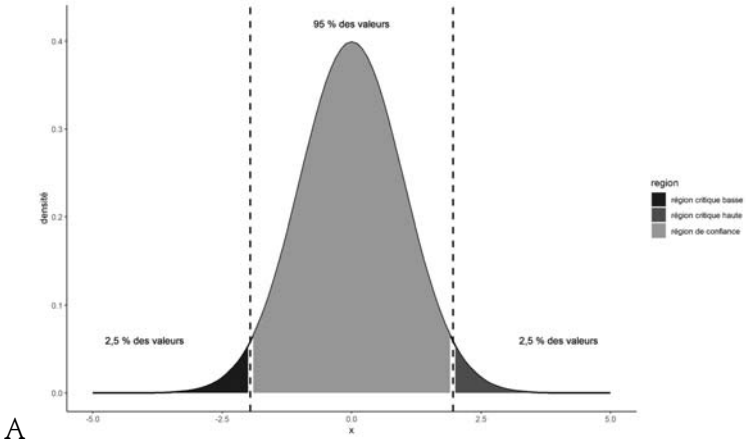
Cas de l'autotest COVID : la sensibilité du test clinique est égale à 0,979 alors que la spécificité est égale à 0,995.

– 5e étape : Conclure en comparant la statistique de test et la valeur critique ou la probabilité calculée et le seuil de significativité (α). Si $p < a$, alors H_0 est rejetée, sinon elle est conservée.

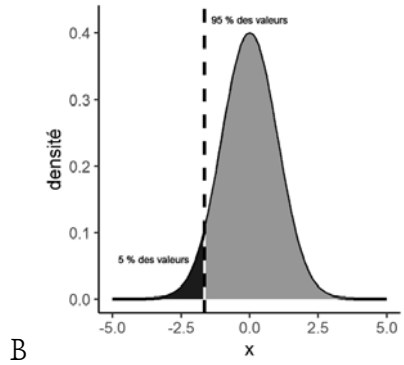
Cas de l'autotest COVID : la sensibilité du test (98 %) est inférieure au seuil de significativité imposé (99 %). L'autotest ne permet pas de détecter de façon fiable l'absence de la COVID.

La statistique de test est calculée en fonction du type et du nombre de données et des hypothèses. Il existe deux grandes catégories de tests : **les tests d'hypothèses paramétriques et non paramétriques**. Les tests d'hypothèses paramétriques s'appuient sur une loi de probabilités, comme la loi normale, la loi de Poisson… ce qui permet une meilleure sensibilité et une plus grande puissance, mais oblige à respecter certaines conditions. Les tests d'hypothèses non paramétriques ne nécessitent pas d'hypothèse sur les lois de probabilités, ils sont donc sont moins restrictifs, mais aussi moins sensibles et moins puissants.

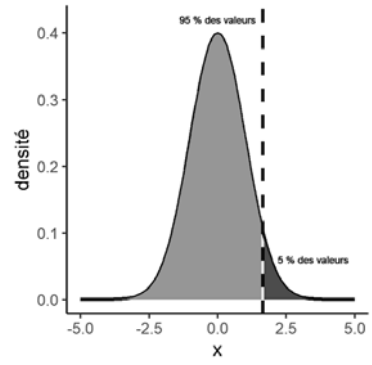
Les données vraies, mais peu courantes, rejetées par le seuil de significativité peuvent être distribuées dans les valeurs extrêmes (test bilatéral, A sur la figure 04-01) ou uniquement dans les valeurs minimales (test unilatéral à gauche, B sur la figure 04-01) ou uniquement dans les valeurs maximales (test unilatéral à droite, C sur la figure 04-01). Les droites verticales en traitillé représentent les valeurs critiques, c'est-à-dire les valeurs déterminées par le seuil de significativité et le type de test.



A



B



C

Figure 04-01 : Test d'hypothèse (ici basé sur une loi normale centrée réduite pour l'exemple) bilatéral, (A), unilatéral à gauche (B) et unilatéral à droite (C). Le seuil de significativité (α) est fixé à 5%.

A. Introduction

Les chapitres précédents nous ont permis de comprendre l'intérêt des data dans la démarche marketing (chapitre Les data, pour quoi faire ?), de découvrir les outils de la data (chapitre Les outils du data marketing) et d'auditer son système data (chapitre Auditer son système marketing). Dans ce quatrième chapitre, nous allons rentrer concrètement dans la manipulation des données. Pour cela, nous allons aborder le premier niveau d'analyse des données, l'analyse descriptive. Nous verrons également certains tests qui relèvent des statistiques inférentielles.

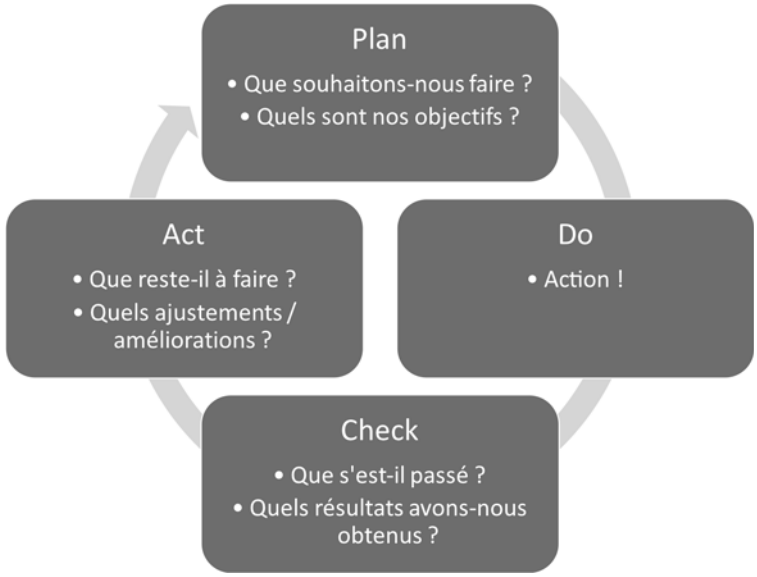
Les statistiques inférentielles permettent, à partir d'un échantillon, d'extrapoler des comportements sur la population totale.

Les outils utilisés seront Excel ou R. Ils nous permettront très simplement d'avoir une première approche des données.

B. Choisir ses KPI

Les KPI (*Key Performance Indicators*) sont les indicateurs qui vont vous permettre de suivre votre activité marketing et commerciale. Ils vous permettent de donner un cap à vos actions (objectifs), de mesurer la réussite des actions lorsqu'elles ont été mises en œuvre et, éventuellement, de réajuster ces actions dans le temps en fonction de l'analyse que vous aurez faite des résultats.

Il s'agit d'une démarche PDCA : `Plan`, `Do`, `Act`, `Check` qui doit guider vos actions et votre stratégie.



Les data sont au service de cette démarche d'amélioration. Elles vont permettre de planifier l'activité et ses objectifs, de mesurer a posteriori les résultats obtenus en les comparant aux objectifs et aux résultats des périodes précédentes. Enfin, elles permettront de réajuster les actions futures.

1. Plan : planifier les actions marketing et les objectifs

Les KPI vont donc être l'outil de pilotage de votre stratégie marketing et de chacune de vos actions. Ils vont également permettre de comparer les résultats entre eux. Par exemple, par canal d'achat, par établissement, par commercial, etc.

Le nombre de KPI pouvant être suivi est infini. Ils dépendent de votre activité (online ou offline, B to C ou B to B, du secteur d'activité), de votre position sur le marché (outsider, suiveur, leader) et de votre stratégie. Il convient de retenir ceux qui vous correspondent le mieux. Il est important d'intégrer l'ensemble de l'équipe dans le choix et la définition des KPI. En effet, si vous avez une vision stratégique, vos collaborateurs ont, eux, une vision plus opérationnelle et pourront vous aiguiller sur la pertinence de tel ou tel indicateur. Également, ils pourront vous indiquer si la donnée correspondante peut réellement être collectée. Enfin, dans la mesure où les KPI peuvent servir de base à l'évaluation de vos équipes, il est préférable de les intégrer en amont dans leur définition pour en faciliter leur compréhension, leur acceptation et l'implication de chacun dans l'atteinte des objectifs.

Vous définirez également la temporalité pertinente pour mesurer chacun de ces indicateurs : par mois, semaine, jour, heure.

2. Do : la phase d'action

La phase d'action est celle où les actions sont mises en œuvre. Il peut s'agir de faire évoluer le site web, d'envoyer une newsletter ou un mailing. Cela peut également être le nombre de leads générés par une nouvelle action marketing.

Cette phase est plus ou moins longue selon la temporalité définie précédemment. Cela peut être une journée, une semaine, un mois ou une année de ventes dans un point de vente.

Si sa durée est longue, il sera préférable d'inscrire des points d'étape pour éventuellement corriger en cours de route les actions.

3. Check : la phase d'analyse et de diagnostic

Une fois la fin de la période définie atteinte, il est temps de passer à l'analyse. Avec les données collectées, vous pourrez procéder au premier niveau d'analyse de vos données : l'analyse descriptive. L'**analyse descriptive** correspond à une phase d'analyse du passé pour mesurer ce qu'il s'est passé et en tirer des enseignements.

Elle est trop souvent négligée au profit de l'action et cette négligence entraîne des ajustements tactiques parfois infondés ou mal orientés. Comme chez le médecin, il faut prendre le temps de discuter, de lire les données (que ce soit un bilan sanguin ou un tableau de bord de résultats) pour poser le bon diagnostic, car c'est ce bon diagnostic qui permettra d'établir la bonne prescription.

Dans cette phase, les données chiffrées seront essentielles, mais il ne faudra pas oublier de discuter avec ceux qui ont participé à l'action pour avoir des feedbacks qualitatifs. C'est votre capacité à confronter ces deux éléments d'analyse qui vous permettra de poser le bon diagnostic.

4. Act : ajuster les actions ou poursuivre sur la même voie

En fonction des résultats obtenus, vous pourrez prendre deux grands types de décisions : ajuster les actions ou poursuivre les actions mises en place. Si des résultats mauvais orientent sur la première voie et des résultats excellents sur la seconde, lorsque les résultats sont mitigés, c'est votre finesse d'analyse qui vous guidera. Plus les résultats sont entre deux eaux, plus il faudra faire de tests pour comprendre en profondeur la situation.

Voyons maintenant concrètement les tests qui peuvent être menés en ce sens.

C. Analyse univariée : étudier les variables une à une

La première étape de l'analyse descriptive consiste à analyser les variables une à une pour avoir une première compréhension des données disponibles. On appelle cette analyse l'analyse univariée en opposition avec les analyses multivariées qui traitent de plusieurs variables à la fois. L'analyse descriptive univariée comprend des mesures souvent connues, mais parfois mal interprétées, nous allons donc passer en revue les indices statistiques présentant les tendances centrales (la moyenne, la médiane, le mode), la dispersion (la variance, l'écart-type, les minimum et maximum, la fréquence et les fractiles) et la forme des données (asymétrie et aplatissement). Enfin, nous aborderons en complément l'intervalle de confiance.

L'objectif ici est de résumer au mieux chacune des variables afin de faciliter la compréhension d'un phénomène (achat de produits, visites en magasin, visites sur un site Internet par exemple). Nous allons donc passer d'une base de données comprenant de nombreuses observations à quelques indicateurs-clés. Pour illustrer ce type d'analyse, nous prendrons l'exemple d'un hôtel qui propose depuis peu ses chambres à différents prix en fonction de la date de réservation et du support de réservation (site Internet, référenceur, etc.).
Il souhaite s'assurer de la rentabilité de cette nouvelle stratégie de prix. La base de données que vous pouvez retrouver dans le classeur **BDD_exemples_chapitre_4.xlsx**, comprend 200 réservations et se présente comme suit :

Client	Canal achat	Type de chambre	Délai réservation (en jours)	Nombre de nuits	Tarif _nuit	CA	satisfaction
927	téléphone	supérieure	4	2	101	202,00	9
549	booking.com	suite familiale	16	2	177	354,00	10
529	mail	suite familiale	6	5	173	865,00	10
885	trivago.fr	standard	6	2	83	166,00	8
883	booking.com	suite familiale	10	4	161	644,00	4
999	booking.com	suite familiale	9	6	170	1020,00	9
960	site internet	standard	8	5	110	550,00	7
419	téléphone	supérieure	8	4	135	540,00	10
434	tripadvisor.fr	standard	7	4	80	320,00	6
517	téléphone	supérieure	5	6	123	738,00	8

À partir de cette base de données, nous allons calculer les indices de tendance centrale et de dispersion.

1. La tendance centrale

Dans le chapitre Auditer son système marketing, nous avions évoqué les différents types de variables : numériques, ordinales ou nominales. Lorsque nous souhaitons connaître la tendance centrale de données, plusieurs indices statistiques existent. Le choix d'un indice dépend du type de variable. Ainsi pour des données numériques (par exemple, le CA) nous pouvons calculer la moyenne, la médiane et le mode. Pour des données ordinales (évaluation de l'image d'une marque), nous pourrons calculer la médiane et le mode. Enfin, pour des données nominales, nous pourrons calculer le mode (canal d'achat, par exemple).

Type de variables	Indice de tendance centrale
Numérique	Moyenne, Médiane, Mode
Ordinales	Médiane, Mode
Nominale	Mode

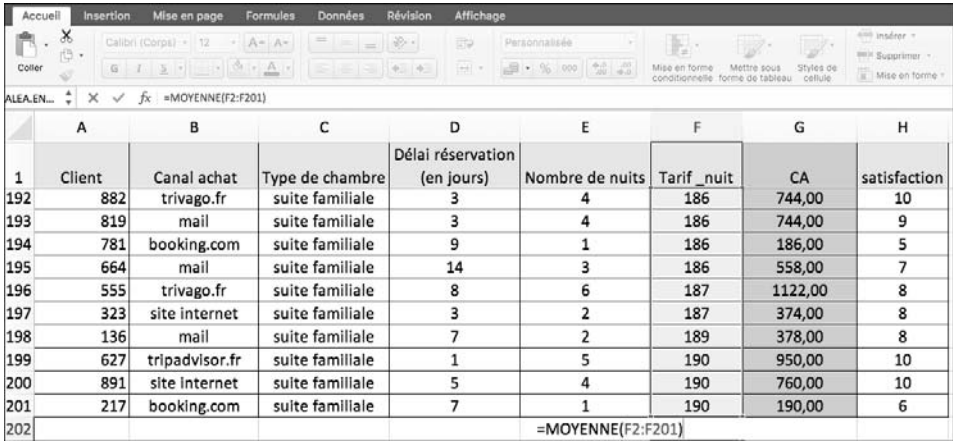
a. La moyenne

La moyenne est un indice de tendance centrale qui permet de résumer les valeurs d'une variable numérique. Dans l'exemple de l'hôtel, nous souhaitons dans un premier temps calculer le prix moyen payé par les clients. La formule de la moyenne est la suivante :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

✎ Dans la cellule F202 d'Excel, saisissez `=moyenne(F2:F201)`.

Cette formule demande à Excel de calculer la moyenne des valeurs allant de la cellule F2 à la cellule F201.



	A	B	C	D	E	F	G	H
1	Client	Canal achat	Type de chambre	Délai réservation (en jours)	Nombre de nuits	Tarif _nuit	CA	satisfaction
192	882	trivago.fr	suite familiale	3	4	186	744,00	10
193	819	mail	suite familiale	3	4	186	744,00	9
194	781	booking.com	suite familiale	9	1	186	186,00	5
195	664	mail	suite familiale	14	3	186	558,00	7
196	555	trivago.fr	suite familiale	8	6	187	1122,00	8
197	323	site internet	suite familiale	3	2	187	374,00	8
198	136	mail	suite familiale	7	2	189	378,00	8
199	627	tripadvisor.fr	suite familiale	1	5	190	950,00	10
200	891	site internet	suite familiale	5	4	190	760,00	10
201	217	booking.com	suite familiale	7	1	190	190,00	6
202					=MOYENNE(F2:F201)			

On obtient pour cette base de données une moyenne de 136,89 €. Nous verrons dans la section Analyse bivariée : faire des rapprochements entre deux variables de ce chapitre comment aller plus loin en comparant les moyennes selon une deuxième variable (canal d'achat, type de chambre, etc.).

b. La médiane

La médiane est un indice de tendance centrale pouvant être utilisée pour les données numériques et ordinales. Elle indique la valeur de la variable qui partage l'échantillon en deux parts égales. Pour la calculer, nous trions les valeurs de la variable par ordre croissant. La valeur partageant l'échantillon en deux parts égales est la médiane.

Si l'échantillon comprend un nombre impair d'observations, elle est facile à repérer. Par exemple, s'il existe 201 clients dans un fichier. Le 101e client (après le tri de la variable en ordre croissant) portera la valeur médiane, car il y aura 100 clients portant une valeur inférieure (clients 1 à 100) et 100 clients portant une valeur supérieure (clients 102 à 201).

Si l'échantillon comprend un nombre pair d'observations, il s'agira alors de toutes valeurs se situant entre les deux valeurs centrales. Ainsi, dans notre exemple de l'hôtel, il y a 200 clients. Comme il s'agit d'un nombre pair, la médiane peut être toute valeur se situant entre celle du client 100 et celle du client 101 (toujours après tri en ordre croissant des valeurs). Pour donner une valeur unique, il est admis d'utiliser la moyenne des valeurs des clients 100 et 101.