



HEC MONTRÉAL

30-650-17

**Introduction à
l'analytique d'affaires**

**Thème 1 : L'analyse
exploratoire des
données**

Les statistiques descriptives

Objectifs de la statistique descriptive

Problématique :

- ❑ Notre cerveau a de la difficulté à assimiler l'information contenue dans une longue liste de nombres. Imaginez une base de données avec des milliers de lignes!

Solution :

- ❑ La statistique descriptive a pour but de résumer l'information contenue dans un ensemble de données sous une forme facilement compréhensible pour notre cerveau humain.

Objectifs de la statistique descriptive

Pour caractériser une distribution de données, en plus de la représentation par des tableaux de fréquences et par des graphiques, on a besoin de connaître:

- ❑ Les points de concentration des données (**mesures de tendance centrale**);
- ❑ Le rang de différentes valeurs, par rapport à l'ensemble des valeurs de la distribution (**mesures de position**);
- ❑ L'importance des écarts observés entre les valeurs (**mesures de dispersion**);
- ❑ Les caractéristiques géométriques de la distribution (**mesures de géométrie**).

Jeux de données

Un **jeu de données** prend souvent la forme d'un tableau où:

- ❑ chaque **colonne** correspond à **une variable** (nom, salaire, âge, etc.);
- ❑ chaque **ligne** contient **les informations** d'un individu ou d'une unité statistique.

Il existe plusieurs **types de variables** qu'il est important de reconnaître. En effet, on ne peut pas traiter toutes les variables de la même façon...

Livres de la série Harry Potter

Voici un petit exemple où 7 lecteurs répondent à des questions à propos de leur tome favori dans la série Harry Potter.

Âge	Genre	HP préféré	Appréciation
10	0	1	4
12	1	3	3
18	0	7	3
14	1	3	4
25	1	5	2
44	1	7	3
8	0	2	4

Les **variables** sont :

- ☐ **Âge** : exprimé en années;
- ☐ **Genre** (sexe) : codé ainsi (0 = homme, 1 = femme);
- ☐ **HP préféré** : le volume, 1 à 7;
- ☐ **Appréciation** (de la série entière) :
1 = détesté, 2 = pas aimé, 3 = aimé, 4 = adoré.

Types de variables

Que peut-on conclure du fait que dans l'échantillon sur les livres de la série *Harry Potter*:

- a) L'âge moyen est 18,7 ans?
 - b) La moyenne de HP préféré est 4 ?
 - c) Le genre moyen est 0,57 ?
-
- Ne vous en faites pas trop! Cela ne signifie pas que les livres d'Harry Potter sont lus par des hermaphrodites qui préfèrent le 4e livre de la série (c'est le plus plate, non?).
 - Même si une variable est codée à l'aide d'un nombre, elle n'est pas nécessairement quantitative.
 - Même si la bourde peut sembler triviale dans cet exemple, ce type d'erreur est très courant, surtout lorsque des non statisticiens utilisent des modèles plus élaborés (e.g. régression).

Types de variables

Qualitative : Décrit des caractéristiques ou des attributs. Elle peut être codée à l'aide de nombres choisis arbitrairement.

- ❑ **Nominale** : Étiquettes associées à chaque catégorie. Ne correspondent pas à des unités de mesure connues et n'impliquent pas un ordre.
Ex: Genre, HP préféré, état civil, une couleur, la marque d'un item.
- ❑ **Ordinale** : Les catégories ont un ordre défini, mais leur valeur numérique n'a pas de signification directe.
Ex: Appréciation, échelles avec des adjectifs, e.g. lente, moyenne, rapide, classe d'âge ou de salaire.

Types de variables

Quantitative (numérique) : Valeurs mesurées sur une échelle.

- ❑ **Discrète** : Les valeurs possibles sont isolées. La plus petite différence entre deux valeurs distinctes est > 0 .

Ex: Nombre de réclamations, nombre d'appels, nombre de *Like*.

- ❑ **Continue** : Les valeurs possibles couvrent un intervalle ou une collection d'intervalles. Même les données arrondies pourraient être traitées comme continues en autant que la plus petite différence entre deux valeurs distinctes est faible.

Ex: Temps de traitement d'une tâche informatique, rendements d'un portefeuille, consommation d'énergie.

Mise en garde

Attention! Le type de variable dépend de la façon dont les données sont présentées ou recueillies.

- ❑ Si on présente le salaire sous forme de classes (ex : moins de 15000\$, entre 15000\$ et 50000\$, 50000\$ et plus), alors cette variable est ordinale.
- ❑ Si on présente la valeur numérique du salaire (en milliers de dollars), cette variable est discrète.
- ❑ Si on présente et dispose du salaire exact, on peut la considérer comme continue même si techniquement on peut aussi dire qu'elle est discrète.

Tableaux croisés

Pour analyser des données, il est très pratique de pouvoir les représenter « conditionnellement ».

Par exemple, dans l'exemple des admissions à Berkeley, il a été utile de voir les taux d'admission par sexe, puis par programme.

Avec Excel, on peut construire de tels tableaux avec la fonctionnalité « **Tableaux croisés dynamiques** ».

Une [capsule vidéo](#) disponible sur zone cours permet d'apprendre comment faire de tels tableaux.

Mise en situation: Cas ProFitness

Un homme d'affaire de renommé vient d'acquérir le centre sportif « ProFitness ».

La seule préoccupation du nouveau propriétaire est de pouvoir préserver la clientèle du centre et éventuellement améliorer la visibilité du centre sportif.

Quelques semaines après l'acquisition, un questionnaire a été envoyé aux 1833 membres du club.

Le sondage contient 9 questions dont 4 spécifiquement conçues pour évaluer le niveau de satisfaction de la clientèle actuelle.

Source: Mise en situation adapté du cas AJFitness tiré du livre « A course in Business Statistics » page 69-70.

Mise en situation: Cas ProFitness

Les variables:

1. **Satisfaction vis-à-vis des équipements**
2. **Satisfaction vis-à-vis du personnel**
3. **Satisfaction vis-à-vis des programmes offerts**
4. **Niveau de satisfaction global**
5. **Nombre d'années d'adhésion**
6. **Sexe** (1 = Homme; 2= Femme)
7. **Nombre de visites par semaine**
8. **Âge en années**
9. **Profession** (1=Étudiant; 2=Professionnel; 3=Cadre; 4=Retraité; 5=Chômeur)

Le niveau de satisfaction a été évalué sur une échelle ordinale codée de 1 à 5.

1	2	3	4	5
Pas satisfait du tout	Insatisfait	Neutre	Satisfait	Très satisfait

Données ProFitness

Le fichier « **1-605-ProFitness_Données.xlsx** » contient les réponses obtenues aux 9 questions du sondage.

Partie 1 :

À l'aide d'Excel, tentons de répondre aux questions suivantes.

1. Calculer le taux de réponse au sondage.
2. Identifier le type de chaque variable dans la base de données.
3. La clientèle du centre (les répondants) est-elle majoritairement masculine ?
4. Quelle est la proportion de Femmes globalement très satisfaites des services offerts par le centre?
5. Commenter le profil des répondants (âge et profession).

Les graphiques

Les graphiques sont un outil puissant pour synthétiser et visualiser les données.

On présentera les graphiques suivants :

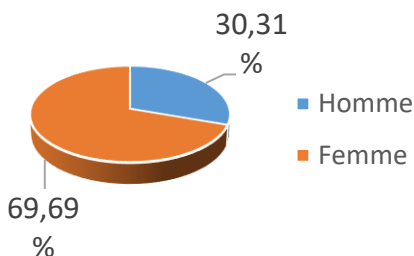
- ☐ Graphique en pointes de tarte,
- ☐ Graphique à bâtonnets,
- ☐ Histogramme.

Graphique en pointes de tarte

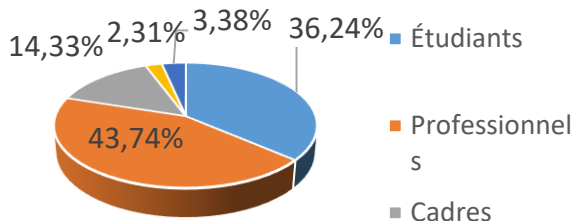
On peut représenter les distributions des variables Sexe et Profession par des pointes en tarte (Diagramme circulaire / pie chart)

- ❑ Les tableaux croisés dynamiques sont utiles pour calculer les chiffres nécessaires.
- ❑ Des [capsules vidéo](#) portant sur la création de graphiques avec Excel sont disponibles. Ne vous fiez pas trop aux recommandations d'Excel.

**Répartition du sexe
chez les répondants**

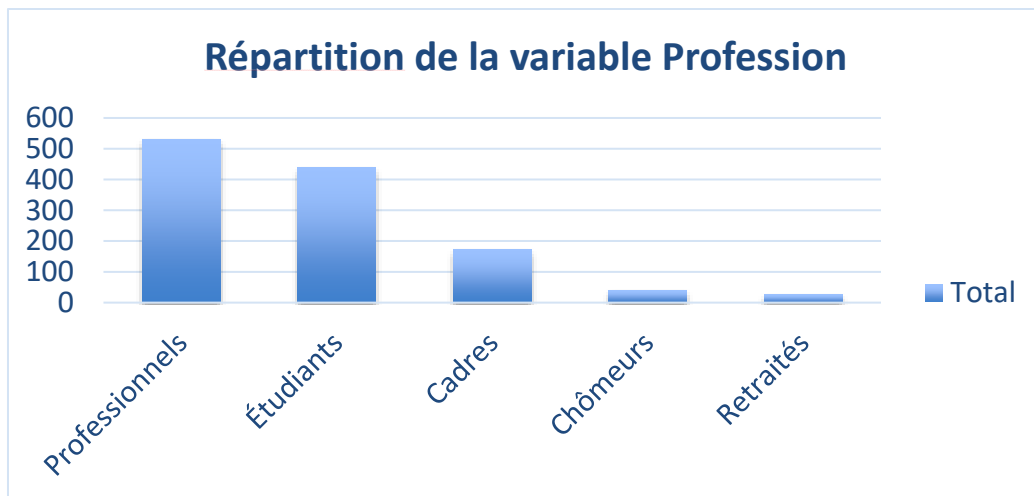


**Répartition de la profession
chez les répondants**



Graphique à bâtonnets

De même, les graphiques à bâtonnets permettent de représenter la répartition d'une variable nominale ou ordinale. Par exemple,



Attention: Même si Excel appelle ces graphiques des histogrammes, ce ne sont pas des histogrammes!

Graphique à bâtonnets

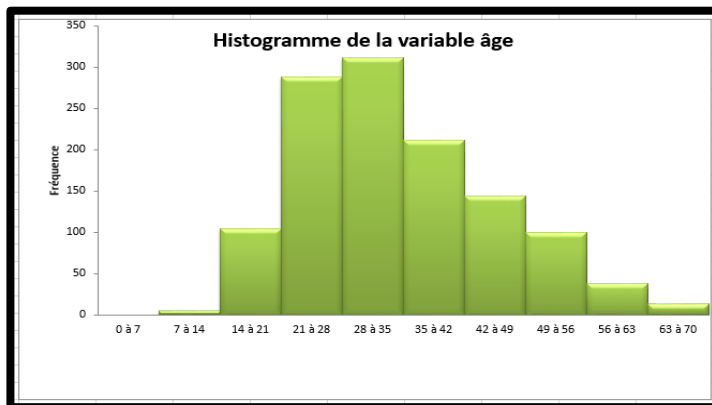
Les graphiques à bâtonnets sont aussi utilisés pour représenter les fréquences d'une **variable quantitative discrète** ayant un nombre fini de valeurs. Par exemple,



Histogramme

Un histogramme est un graphique représentant la distribution d'une **variable quantitative continue**.

La variable sur l'abscisse (axe des x) doit être une variable continue qui est **découpée en intervalles**. Par exemple, voici la distribution de l'âge des répondants du centre :



Chaque colonne correspond à un intervalle et sa taille est proportionnelle à la fréquence (le nombre répondant dans ce cas-ci).

Histogramme

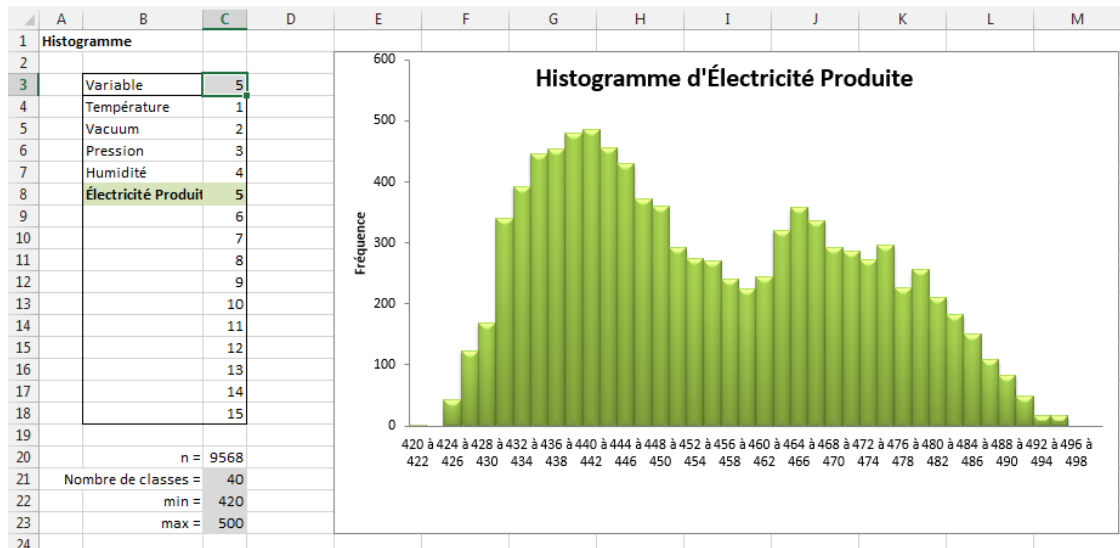
Créer un histogramme n'est pas simple avec Excel. Il faudrait d'abord créer les catégories et calculer les fréquences avant de produire le graphique à bâtonnets.

Heureusement, nous avons préparé un gabarit Excel permettant de créer automatiquement un histogramme.

Le gabarit est disponible sur Zone Cours. « **StatistiquesDescriptives.xlsx** » (maximum 10 000 données).

Une [capsule vidéo](#) disponible sur Zone Cours explique le fonctionnement de ce gabarit.

Présentation brève du gabarit



Quel graphique choisir?

- ☐ **Variable numérique** : histogramme ou diagramme en bâtonnets pour une variable discrète ayant un nombre fini de valeurs.
- ☐ **Variable nominale** :
 - ☐ Fréquence : Pointe de tarte ou graphique à bâtonnets.
 - ☐ Résumé d'une variable numérique pour chaque valeur de la variable nominale : graphique à bâtonnets.

Si une variable numérique discrète présente peu de valeurs possibles, elle peut aussi être représentée par un graphique à pointe de tarte (peu recommandé).

La Presse+, 18 mai 2014



« Nous constatons aussi que les gens sous-estiment les risques de longévité. » — Hélène Gagné, gestionnaire de portefeuille chez Gestion privée Peak.

Pourtant, comme l'espérance de vie représente la durée de vie **moyenne**, « il y a **une chance sur deux (50 %)** pour que nous la dépassions ».

Indicateurs numériques

On se sert souvent d'indicateurs numériques pour décrire une population ou un échantillon.

Exemples d'indicateurs numériques :

- ☐ Moyenne
- ☐ Médiane
- ☐ Proportion
- ☐ Écart-type
- ☐ Centile

Toutefois ces indicateurs ne sont pas toujours bien compris...

Discussion

- ❑ Pourtant, comme l'espérance de vie représente la durée de vie **moyenne**, « il y a **une chance sur deux (50 %) pour que nous la dépassions** ». Il y a une chance sur 2 de dépasser la moyenne?
- ❑ Un politicien promet que tout le monde gagnera au-dessus de la moyenne. Est-ce possible?
- ❑ Un professeur dit que son cours est tellement facile que tous les étudiants auront une note au-dessus de la moyenne...?

Mesures de centralité : la moyenne

Formule :

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

Propriétés :

- ☐ Mesure de centralité la plus utilisée
- ☐ Souvent mal interprétée
- ☐ Seulement utilisable pour des variables quantitatives.
- ☐ Influencée par les valeurs extrêmes : une seule donnée très grande (ou très petite) peut occasionner un grand changement dans sa valeur (**effet Bill Gates**).
- ☐ Lorsque les valeurs prennent les valeurs 0 ou 1, la moyenne se nomme alors **proportion**.
- ☐ Si l'on répète une expérience aléatoire un grand nombre de fois, la moyenne des résultats devient presque constante.

Quel est le salaire typique d'un Québécois?

Quel est le salaire typique d'un Québécois?

Publié le 04 septembre 2013 à 06h40 | Mis à jour le 04 septembre 2013 à 06h40



FRANCIS VAILLES
La Presse



La question est toute simple, mais la réponse est plus compliquée qu'il n'y paraît: quel est le salaire typique d'un Québécois, avant impôts?

Mercredi dernier (28 août), justement, Statistique Canada a sorti son bulletin mensuel, repris par plusieurs médias. Selon l'organisme, la rémunération moyenne s'élevait à 919\$ en juin, en hausse de 0,2% par rapport au mois précédent. Extrapolée sur une année, la rémunération moyenne friserait donc les 48 000\$, ce qui semble plus élevé que la réalité, effectivement.

Première observation: il s'agit d'une moyenne canadienne. Les médias ne donnent pas les chiffres par province, s'intéressant davantage à la hausse (0,2%) qu'au chiffre absolu. Vérification faite, pour le Québec, cette rémunération était de 838\$, en baisse de 0,2%, soit environ 43 500\$ par année.

Fin de la discussion? Pas vraiment. D'abord, le salaire moyen ne traduit pas nécessairement bien la réalité. En effet, un échantillon de deux individus où le premier gagnerait par exemple 5000\$ par mois et le second, 500\$ nous donnerait une moyenne de 2750\$, ce qui ne représenterait pas fidèlement la réalité.

Pour contourner ce problème, un des indicateurs souvent utilisés par les économistes est la médiane, soit la valeur qui permet de séparer un groupe exactement en deux. Par exemple, dans un groupe de 100 employés, le salaire médian serait celui du 50e mieux payé.

Selon cette enquête, donc, les travailleurs québécois à temps plein ont déclaré avoir gagné 896\$ en juin, en moyenne, pour une semaine de travail de 38 heures (23,57\$ de l'heure). La médiane est de 800\$ (21\$ de l'heure). La moyenne canadienne est 12% plus élevée et la médiane, 10%.

En somme, selon les enquêtes, le salaire typique (ou médian) des Québécois oscille probablement entre 800\$ et 837\$ par semaine, soit l'équivalent de 41 600\$ à 43 500\$ par année.

896 \$

Salaire moyen à temps plein

789 \$

Salaire moyen à temps plein et partiel

800 \$

Salaire médian des temps pleins

717 \$

Salaire médian des temps pleins et partiels

Mesures de centralité: la médiane

La **médiane** est la valeur de la variable qui départage les données en deux parties égales (en moitié). Elle indique **le centre des données ordonnées**.

Exemple 1: Calcul de la médiane des valeurs 2, 19, 5, 12, 2

2 2 5 12 19 → médiane = 5

Si n est impair, alors la médiane est la $\left(\frac{n+1}{2}\right)^{\text{ème}}$ donnée.

Exemple 2: Calcul de la médiane des valeurs 2, 9, 19, 5, 12, 2

2 2 5 9 12 19 → médiane = $(5+9)/2=7$

Si n est pair, alors la médiane est le point milieu des $\left(\frac{n}{2}\right)^{\text{ème}}$ et $\left(\frac{n}{2} + 1\right)^{\text{ème}}$ données.

On peut dire que 50% des observations sont plus petites ou égales à 7 ou 50% des observations sont plus grandes ou égales à 7 dans l'exemple 2.

Mesures de centralité : la médiane

Propriétés :

- ☐ Mesure de localisation moins utilisée et moins connue que la moyenne.
- ☐ La médiane est plus robuste que la moyenne (peu ou pas influencée par les valeurs extrêmes).
- ☐ Facile à interpréter.

« Nous constatons aussi que les gens sous-estiment les risques de longévité. »

— Hélène Gagné, gestionnaire de portefeuille chez Gestion privée Peak

Pourtant, comme l'espérance de vie représente la durée de vie **moyenne**, « il y a **une chance sur deux (50 %) pour que nous la dépassions** ».

La gestionnaire mélangeait donc de la durée de vie moyenne et la durée de vie médiane!

Effet Bill Gates

Trois clients sont dans un bar. Leurs fortunes respectives sont 650 000 \$, 450 000 \$, et 700 000 \$. La fortune moyenne est donc 600 000 \$ et la fortune médiane est 650 000 \$.

Bill Gates arrive dans le bar. Sa fortune personnelle est estimée à 76 milliards.

☐ Quelle est la nouvelle moyenne? **Rép: 19 milliards!**

⇒ La moyenne est très affectée par l'arrivée de Bill Gates.

☐ Quelle est la nouvelle médiane? **Rép: 675 000 \$.**

⇒ La médiane est très peu affectée par l'arrivée de Bill Gates.

Moyenne VS Médiane

Le tableau suivant résume les dépenses annuelles moyennes des joueurs en ligne, a-t-on raison de s'alarmer?

Tableau 1b.
Dépenses annuelles totales (\$CAD) des différents types de joueurs et selon le lieu de jeu, Québec, 2009

	N	Moyenne	Médiane	Moyenne excluant les valeurs aberrantes	% de valeurs aberrantes	Intervalle des valeurs aberrantes
Joueurs en ligne						
Oui	111	9 904	856	2 300	7,21 %	21 305 – 306 270
Non	8 070	527	128	406	0,33 %	15 164 – 114 082
Joueurs au casino						
Oui	1 057	2 381	349	950	1,99 %	16 940 – 306 270
Non	7 119	432	120	357	0,20 %	15 164 – 114 082
Joueurs d'ALV						
Oui	432	3 972	767	1 409	6,02 %	15 164 – 114 082
Non	7 744	483	120	376	0,12 %	18 724 – 306 270

Source : https://www.concordia.ca/content/dam/artsci/research/lifestyle-addiction/docs/projects/enhjeu-q/ENHJEU-QC-2009_rapport-final-FRQ-SC.pdf

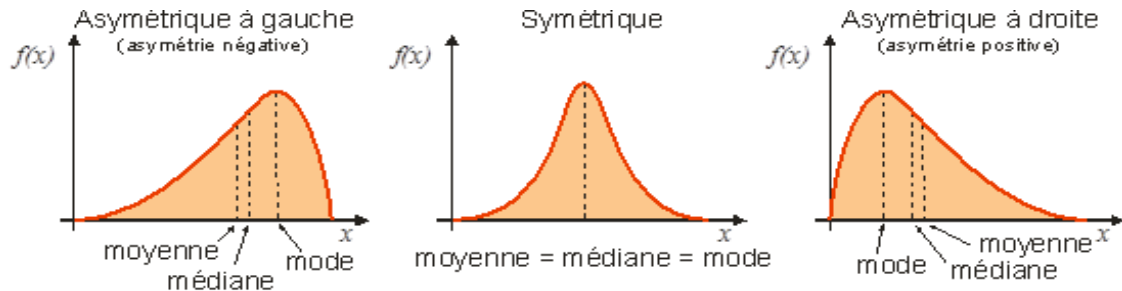
Mesures de centralité : le mode

Le mode :

- ☐ Seule mesure de tendance centrale pour **les données qualitatives**.
- ☐ Le mode est la valeur la plus fréquente.
- ☐ Pour une variable continue, les données doivent être groupées et le mode sera le point milieu de l'intervalle définissant la catégorie avec le plus grand effectif.
- ☐ Le mode est ce qui est le plus à la mode...!
- ☐ **Exemple** : Calcul du mode des valeurs 2, 19, 5, 12, 2 :
 - 2 2 5 12 19 → mode = 2

Positions relatives de la moyenne, de la médiane et du mode

La position relative de la moyenne, de la médiane et du mode dépend de la forme de la distribution (forme de l'histogramme).



- ☐ Pour une distribution **unimodale** symétrique :
mode = médiane = moyenne
- ☐ Pour une distribution **unimodale** asymétrique à droite :
mode < médiane < moyenne
- ☐ Pour une distribution **unimodale** asymétrique à gauche :
mode > médiane > moyenne

Calculs des mesures de centralité avec Excel

- ❑ **Moyenne:** Utiliser la fonction Excel
= **MOYENNE** (*plage de données*)
- ❑ **Médiane:** Utiliser la fonction Excel
= **MEDIANE**(*plage de données*)
- ❑ **Mode:** Utiliser la fonction Excel :
= **MODE**(*plage de données*)

*Remarque : cette formule ne doit être utilisée que pour **des variables discrètes, nominales ou ordinales**.*

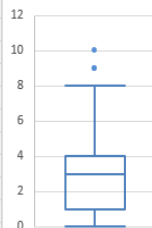
La moyenne et la médiane sont également calculées dans le gabarit Excel **StatistiquesDescriptives.xlsx** (voir la diapo suivante).

Gabarit StatistiquesDescriptives.xlsx

Dans l'onglet **Statistiques** de l'outil « **StatistiquesDescriptives.xlsx** », on retrouve les différents indicateurs. Illustration pour les statistiques sur le nombre de visites par semaine des répondants au sondage:

Statistiques descriptives					
Variable	7	Taille d'échantillon	1214		
Satisfaction vis-à-	1				
Satisfaction vis-à-	2	Mesures de centralité		Quantiles	
Satisfaction vis-à-	3	Moyenne	2.72405272	Min	0
Niveau de satisfa	4	Médiane	3	5%	0
Nombre d'années	5			25%	1
sexe	6	Mesures de dispersion		50%	3
Nombre de visite	7	Écart-type	1.70846244	75%	4
âge	8	Étendue	10	95%	6
visites homme	9	Étendue inter-quartile	3	Max	10
visites femme	10				
	11	Mesures de géométrie		70%	3
	12	Asymétrie	0.63976925		
	13	Aplatissement	0.26434918		
	14				
	15	Probabilités empiriques			
		Valeur d'intérêt :	2		
		P(Nombre de visites par semaine < 2) =	0.27018122		
		P(Nombre de visites par semaine <= 2) =	0.49423394		
		P(Nombre de visites par semaine > 2) =	0.50576606		
		P(Nombre de visites par semaine >= 2) =	0.72981878		

Boxplot de
Nombre de
visites par
semaine



Mesures de dispersion

Étendue= MAX-MIN

L'étendu de la plage couverte par les données donne un aperçu de la dispersion des données mais reste non informative par rapport à ce qui se passe entre les valeurs minimale et maximale.

Variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{x})^2$$

Écart-type

$s = \sqrt{s^2}$: représente l'écart moyen observé entre les données et leur moyenne.

- ☐ L'écart-type partage les mêmes unités que la variable originale.
- ☐ L'écart-type est influencé par les valeurs extrêmes (comme la moyenne).
- ☐ L'écart-type et la variance sont toujours positifs. Plus l'écart-type est faible, plus les données sont concentrées autour de la moyenne.
- ☐ En finance, on mesure la volatilité par l'écart-type. C'est aussi vu comme une mesure de risque.

Calcul des mesures de dispersion avec Excel

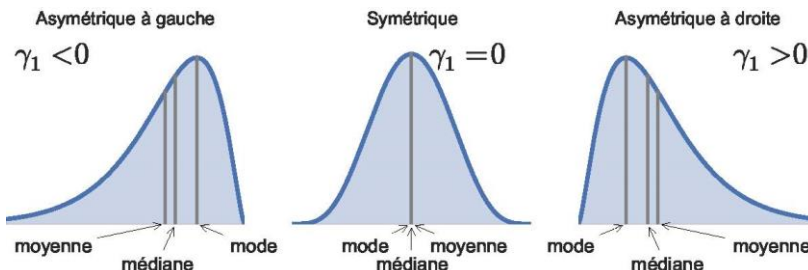
- ❑ **Étendue** : Pas de fonction Excel pré-programmée, écrire la formule :
= **MAX** (plage de données) – **MIN** (plage de données)
- ❑ **Écart-type**: Utiliser la fonction Excel :
= **ECARTYPE.STANDARD** (plage de données)

Note : ces mesures sont également calculées dans le gabarit Excel « *StatistiquesDescriptives.xlsx*. »

Mesure de géométrie : coefficient d'asymétrie

Formule :

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n (x_i - \bar{x})^3 / s^3$$



Coefficient d'asymétrie	Distribution
$= 0$	Parfaitement symétrique
< 0	Asymétrique à gauche
> 0	Asymétrique à droite

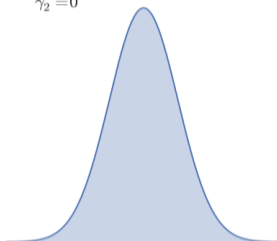
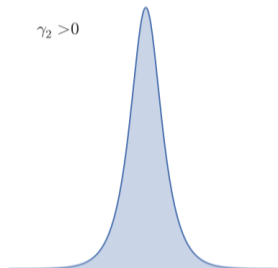
Mesures de géométrie: coefficient d'aplatissement (kurtosis)

Formule :
$$g_2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^4 / s^4 - 3$$

Aplatissement négatif

Aplatissement nul

Aplatissement positif

 $\gamma_2 < 0$  $\gamma_2 = 0$  $\gamma_2 > 0$ 

Coefficient d'aplatissement	Sommet
= 0	Comme la loi Normale
< 0	Plus aplati que la loi Normale
> 0	Plus aigu que la loi Normale

Mesures de géométrie

Propriétés :

- ☐ Ces coefficients sont surtout utilisés en économie et finance.
- ☐ Ne dépendent pas des unités.
- ☐ Sensibles aussi aux valeurs extrêmes.

Comment obtenir les coefficients d'asymétrie et d'aplatissement dans Excel?

- ☐ **Coefficient d'asymétrie:**
Utiliser la fonction = **COEFFICIENT.ASYMETRIE** (*plage des données*)
- ☐ **Coefficient d'aplatissement :**
Utiliser la fonction = **KURTOSIS** (*plage des données*)

On peut aussi obtenir ces indicateurs Avec l'outil « **StatistiquesDescriptives.xlsx** »

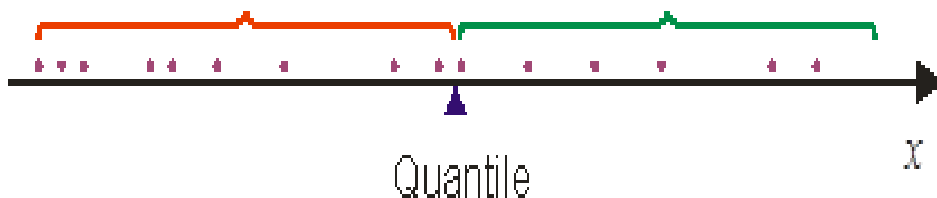
Mesures de position: Les quantiles

Un **quantile d'ordre α** , noté Q_α ($0 < \alpha < 1$) est une valeur telle que la proportion des observations qui lui sont inférieures est tout au plus de α et la proportion des observations qui lui sont supérieures est tout au plus de $(1 - \alpha)$.

$$P(X < Q_\alpha) \leq \alpha \text{ et } P(X > Q_\alpha) \leq 1 - \alpha.$$

Au plus $100\alpha\%$ des données
sont inférieures au quantile

Au plus $100(1-\alpha)\%$ des données
sont supérieures au quantile



Mesures de position: Les quantiles

Exemples:

- ☐ **Médiane** (50%-50%): 2 parties égales
- ☐ **Quartiles** (Q_1 (25%), Q_2 =Md (50%) et Q_3 (75%)): 4 parties égales
- ☐ **Quintiles** (20%-40%-60%-80%) :5 parties égales
- ☐ **Déciles** (10%-20%-30%-40%-50%-60%-70%-80%-90%): 10 parties égales
- ☐ **Centiles** (1%-2%-.....-99%): 100 parties égales


Propriété

- ☐ Assez peu sensibles aux valeurs extrêmes.

Quantiles avec Excel

- ❑ La fonction **centile** (percentile en anglais) fournit le quantile d'ordre **k** d'un échantillon.
- ❑ **Utilisation:** = **centile** (*plage de données, k*)
 - **k** doit être une fraction entre 0 et 1.
 - Par exemple, $k = 0.5$ donne la médiane.
 - Pour le premier quartile, $k = 0.25$.
 - Pour le troisième quartile, $k = 0.75$.
- ❑ **Note** : certains quantiles sont pré-calculés dans le gabarit Excel **StatistiquesDescriptives.xlsx**

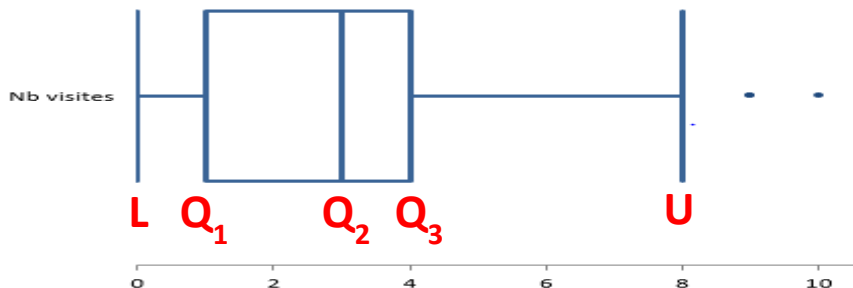
Pour tout autre quantile, indiquer l'ordre k du quantile dans la cellule suivante:



Min	0
5%	0
25%	1
50%	3
75%	4
95%	6
Max	10
70%	3

Boxplot (diagramme en boîte ou boîte à moustaches)

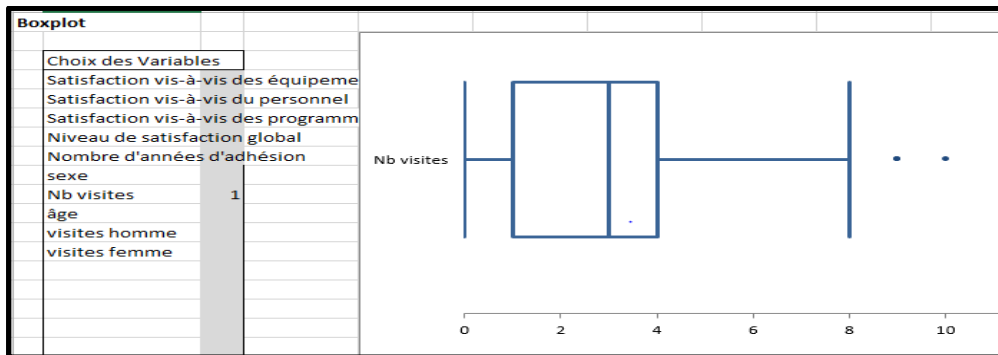
- Le **boxplot** permet de rapidement résumer la position des données.



- La boîte au centre s'étend de Q_1 à Q_3 . Sa longueur est l'intervalle inter-quartile (IQR).
- La ligne au centre de la boîte est Q_2 , la médiane.
- Les moustaches (lignes horizontales) joignent la boîte aux barres verticales (situées en L et U) délimitant les frontières avec les valeurs extrêmes. Chaque valeur extrême (à l'extérieur de la région délimitée par les barres verticales) est marquée par un point.
- La **symétrie du diagramme en boîte est associée à la symétrie de la distribution.**

Comment construire un boxplot?

Le gabarit **StatistiquesDescriptives.xlsx** offre également un boxplot des données dans l'onglet « Boxplot ».



Voir les [capsules-vidéo](#) disponibles dans zone cours.

Il est possible d'obtenir plusieurs Boxplots dans le même graphique en marquant 1 pour les variables d'intérêt.

Résumé : indicateurs numériques

	Mesure	Qualitative		Quantitative	
		Nominale	Ordinale	Discrète	Continue
Tendance centrale	Moyenne			✓	✓
	Médiane		✓	✓	✓
	Mode	✓	✓	✓	Groupées
Dispersion	Variance /Écart-type			✓	✓
	Étendue			✓	✓
	Intervalle inter-quartile			✓	✓
Position	Quantiles		✓	✓	✓

Cas ProFitness

Partie 2 :

On s'intéresse à la comparaison du nombre de visites des hommes et des femmes.

1. Calculer la moyenne, la médiane, le coefficient d'asymétrie, l'étendue, l'écart-type et l'IQR pour chaque variable.
2. Comparer la distribution de la variable nombre de visites chez les Hommes et chez les Femmes et commenter (utilisez les Boxplots).
3. Tracer les histogrammes pour chaque variable et commenter.
4. Le nouveau propriétaire désire augmenter le nombre de visites au centre sportif. Quelles recommandations pouvez-vous lui suggérer à la lumière de ces résultats?
5. Le nouveau propriétaire désire améliorer le niveau de satisfaction global. Fiez-vous aux données pour lui communiquer des pistes d'amélioration.