

T. D. n° 1 4056

Résumé

Ce document est le T. D. n° 1 du module 4056. Il reprend rapidement des éléments du cours et propose une mise en pratique interactive des notions de densité, de variance et de normalité.

1 Jeux vidéo

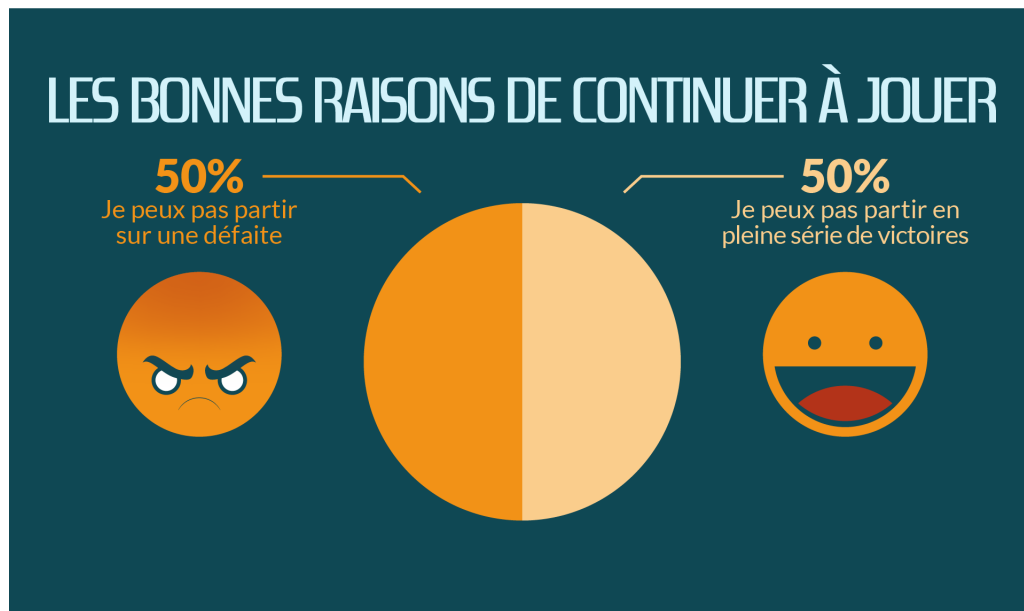


FIGURE 1 – Statistique et jeux vidéo
source : www.c-fun.fr

1.1 Chargement des données

Vous êtes un jeune entrepreneur désireux de faire fortune dans les jeux vidéo. Comme vous avez beaucoup d'imagination mais peu d'argent, vous souhaitez investir là où vous êtes sûr de faire du profit. Vous allez observer des données issues de ventes de jeux vidéo afin de faire votre propre étude de marché.

Cet ensemble de données contient une liste de jeux vidéo avec des ventes supérieures à 100 000 exemplaires. Il a été généré par le site de vgchartz.com.

Les champs incluent

- *Rank* : Classement des ventes globales
- *Name* : Nom du jeu

- *Platform* : Plate-forme de la version des jeux (c'est-à-dire PC, PS4, etc.)
- *Year* : Année de sortie du jeu
- *Genre* : Genre du jeu
- *Publisher* : Éditeur du jeu
- *NA_Sales* : Ventes en Amérique du Nord (en millions)
- *EU_Sales* : Ventes en Europe (en millions)
- *JP_Sales* : Ventes au Japon (en millions)
- *Other_Sales* : Ventes dans le reste du monde (en millions)
- *Global_Sales* : Total des ventes mondiales.

Après avoir téléchargé le fichier *vgsales.csv* chargez-le sous **R** (vous stockerez ces données sous la variable *Data*). Il est également possible de télécharger le fichier *.csv* depuis : <https://www.kaggle.com/gregorut/videogamesales>.

```
>Data <- read.csv("C:/Users/claey/Documents/Cours/cour ESIEA/4a/
  vgsales.csv", sep=',')
```

La fonction **read.csv()** lit des données de type *csv* pour les stocker dans un *dataframe*.

À moins de vouloir faire frémir un statisticien, vous ne touchez pas aux données originales (car vous risquez de faire n'importe quoi au cours de ce T. D. sur ces pauvres données qui n'ont rien demandé). Il faut donc copier les données dans un deuxième *dataframe*.

```
>df<-as.data.frame(Data)
```

On nottera cette variable *df* pour *dataframe*. *Data* était déjà un *dataframe*, mais c'est pour vous montrer la fonction

as.data.frame() qui vous sera très utile par la suite.

Il est temps de distinguer une tendance sur les données! Pour cela, la fonction **summary()** vous donne des des informations « de type position » sur toutes les variables.

À vous !

- a) Cherchez et citez à quoi correspond un objet de type "*dataframe*" sous **R**.
- b) Appliquez la fonction **summary()** sur votre *dataframe*.
- c) Vérifiez la moyenne de la colonne *JP_Sales* avec la fonction **mean()**. Comparez avec celle donnée par la fonction **summary()**.
- d) A priori, quel genre a été le plus vendu ?
- e) A priori, quel pays a eu les meilleures ventes ?
- f) Observez la variable explicative *Year*. Pourquoi cette variable peut contredire les deux hypothèses précédentes ?

```
> summary(df)
      Rank                                     Name
Min.   :      1   Need for Speed: Most Wanted:   12
1st Qu.: 4151   FIFA 14                        :      9
```

Median : 8300	LEGO Marvel Super Heroes	:	9
Mean : 8301	Madden NFL 07	:	9
3rd Qu.:12450	Ratatouille	:	9
Max. :16600	Angry Birds Star Wars	:	8
	(Other)	:	16542
Platform	Year	Genre	
DS :2163	2009 :1431	Action :3316	
PS2 :2161	2008 :1428	Sports :2346	
PS3 :1329	2010 :1259	Misc :1739	
Wii :1325	2007 :1202	Role-Playing:1488	
X360 :1265	2011 :1139	Shooter :1310	
PSP :1213	2006 :1008	Adventure :1286	
(Other):7142	(Other):9131	(Other) :5113	
	Publisher	NA_Sales	
Electronic Arts	: 1351	Min. : 0.0000	
Activision	: 975	1st Qu.: 0.0000	
Namco Bandai Games	: 932	Median : 0.0800	
Ubisoft	: 921	Mean : 0.2647	
Konami Digital Entertainment:	832	3rd Qu.: 0.2400	
THQ	: 715	Max. :41.4900	
(Other)	:10872		
EU_Sales	JP_Sales	Other_Sales	
Min. : 0.0000	Min. : 0.00000	Min. : 0.00000	
1st Qu.: 0.0000	1st Qu.: 0.00000	1st Qu.: 0.00000	
Median : 0.0200	Median : 0.00000	Median : 0.01000	
Mean : 0.1467	Mean : 0.07778	Mean : 0.04806	
3rd Qu.: 0.1100	3rd Qu.: 0.04000	3rd Qu.: 0.04000	
Max. :29.0200	Max. :10.22000	Max. :10.57000	
Global_Sales			
Min. : 0.0100			
1st Qu.: 0.0600			
Median : 0.1700			
Mean : 0.5374			
3rd Qu.: 0.4700			
Max. :82.7400			

1.2 La notion de densité

Vous allez observer la densité de probabilité des données. En physique, la densité ou densité d'un corps est le rapport de sa masse volumique à la masse volumique d'un corps pris comme référence, c'est à dire que nous comparons le nombre d'une variable par rapport à un ensemble de variables observées. En théorie des probabilités ou en statistique, une densité de probabilité est une fonction qui permet de représenter une loi de probabilité sous forme d'intégrales. Dans un histogramme, la densité en un point x est estimée par la proportion d'observations x_1, x_2, \dots, x_N qui se trouvent à proximité de x . Pour cela, nous traçons une boîte en x et dont la largeur est définie par un paramètre de lissage h (soit la largeur de la boîte); nous comptons ensuite

le nombre d'observations qui appartiennent à cette boîte.¹

Le problème avec les histogrammes, c'est que :

- nous devons définir le paramètre h (dans R il est calculé automatiquement)
- les histogrammes produisent une estimation de la fréquence non continue.

La fonction **density()** fournit une estimation par noyau (ou encore méthode de Parzen-Rosenblatt, 1962). C'est une méthode non paramétrique d'estimation de la densité de probabilité d'une variable aléatoire. Elle se base sur un échantillon d'une population et permet d'estimer la densité de probabilité en tout point du support (intervalle min et max des valeurs observées). Elle est plus précise qu'un simple histogramme (fonction **hist()**).

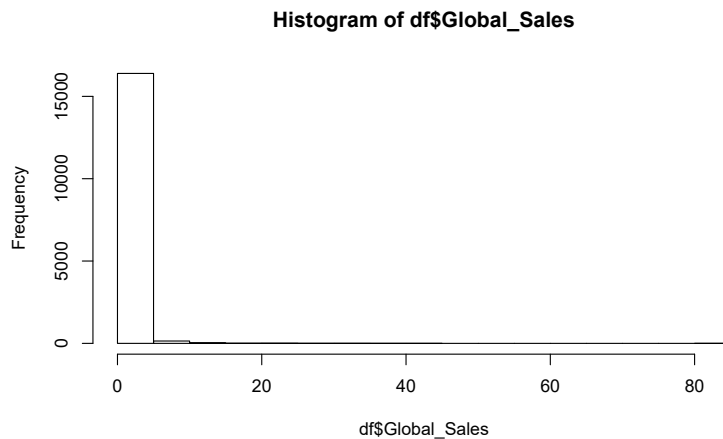


FIGURE 2 – Histogramme des valeurs de *Global_Sales*

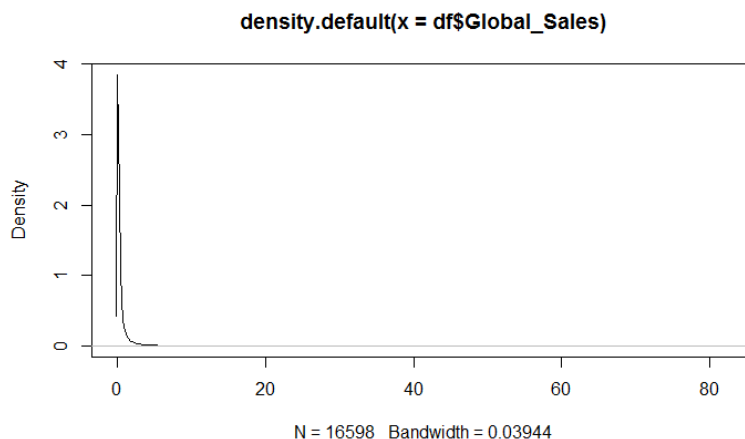


FIGURE 3 – Densité de probabilité des valeurs de *Global_Sales*

1. Ces définitions sont inspirées de celles que vous trouvez sur Wikipédia, vous pourrez ainsi aisément retrouver la définition et la démonstration mathématique.

```
> hist(df$Global_Sales)

> plot(density(df$Global_Sales))
```

La Figure 3 n'est pas très explicite, essayons de savoir pourquoi. Affichez les *Global_Sales* pour chaque année à l'aide de la fonction **plot()**.

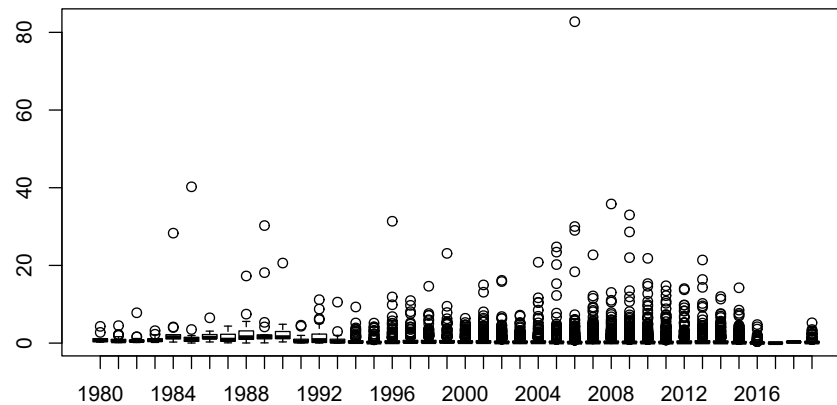


FIGURE 4 – Ventes globales par année

```
> plot(df$Year, df$Global_Sales)
```

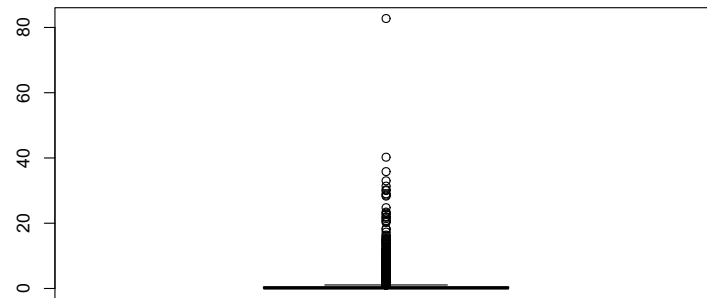
À vous !

- Y-a-il une année qui se démarque des autres ? Laquelle ?
- Pourquoi certaines valeurs extrêmes peuvent-elles fausser l'interprétation d'une moyenne ?

La méthode du noyau consiste à retrouver la continuité : pour cela, nous remplaçons la boîte centrée en x et de largeur h par une loi gaussienne (définie par la suite) centrée en x . Plus une observation est proche du point de support x plus la courbe en cloche lui donnera une valeur numérique importante. À l'inverse, les observations trop éloignées de x se voient affecter d'une valeur numérique négligeable. Notez également que plus il y a d'observations dans le voisinage d'un point, plus sa densité est élevée.

1.3 La variance

Utilisez la fonction **boxplot()** sur votre dataframe *df*.

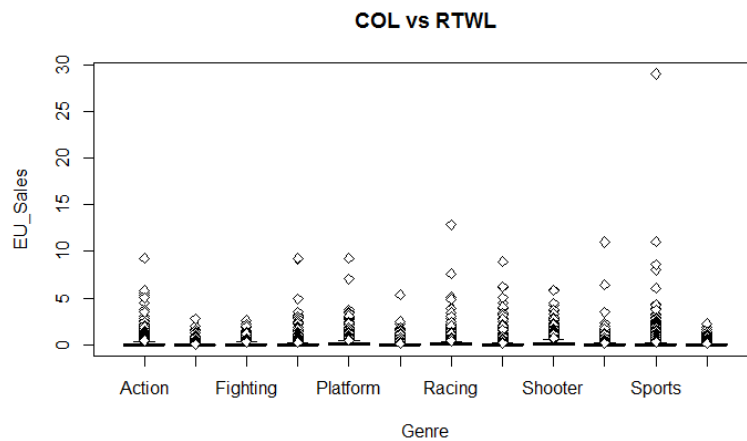
FIGURE 5 – Boxplot des *Global_Sales*

```
> boxplot(df$Global_Sales)
```

À vous !

- Expliquez le principe de la fonction **boxplot()**.
- Expliquez en quoi une valeur perturbe potentiellement notre interprétation du résultat de **boxplot()**.

Nous allons observer les ventes européennes selon les différents genres

FIGURE 6 – Boxplot de *EU_Sales* par genre

```
> boxplot(EU_Sales ~ Genre, col='yellow',
  pch=23,
  xlab = "Genre", ylab = "EU_Sales",
  main='COL vs RTWL',
  data=df)
```

À vous !

- Quel est le genre qui semble avoir eu les meilleures ventes ?
- Affichez la boîte à moustaches des *Global_Sales* en fonction du genre de jeux
- Comparez les deux résultats et concluez.

Il est possible de savoir si un genre de jeux se vend aussi bien qu'un autre. Nous pouvons par exemple observer la variance. La variance est une mesure servant à caractériser la dispersion d'une série de mesures ou d'une distribution. La variance correspond au carré de l'écart-type noté σ . Elle est généralement notée σ^2 ou $\text{Var}(X)$. La variance indique de quelle manière la série de mesures ou la distribution se disperse autour de sa moyenne ou son espérance. Une variance élevée indique que les valeurs sont très écartées les unes des autres et vice-versa. Elle est nulle lorsque toutes les valeurs sont identiques. Remarque : une variance n'est donc jamais négative.

À vous !

- À quel mathématicien devons-nous la découverte de la variance ?
- La variance permet d'obtenir l'écart type, qui est la racine carrée de la variance. Pourquoi l'écart-type est souvent plus parlant que la variance pour appréhender la dispersion ?
- La variance est un des éléments permettant de caractériser une loi de probabilité. Pourquoi ?

Vous décidez de comparer deux genres différents : les jeux d'action et les jeux de sports.

```
> COL0 = df$Global_Sales[df$Genre == 'Action']
> COL1 = df$Global_Sales[df$Genre == 'Sports']
> var(COL0)
[1] 1.337324
```

À vous !

- Calculer avec **R** la variance des jeux de type sport. Quelle est cette valeur ?
- Commentez la variance du jeu *Action* et du jeu *Sports*. Qu'en concluez vous ?

L'analyse de la variance permet d'étudier par exemple le comportement d'une variable qualitative à expliquer en fonction d'une ou de plusieurs variables nominales catégorielles. Cependant, certains tests sont applicables uniquement si les données suivent une loi normale. Il existe des tests statistiques permettant de savoir si une distribution suit la loi normale.

1.4 Un modèle dit "gaussien"

En théorie des probabilités et en statistique, la loi normale est l'une des lois de probabilités les plus adaptées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires. Elle est en lien avec de nombreux objets mathématiques

dont le mouvement brownien, le bruit blanc gaussien pour ne citer qu'eux. Elle est également appelée loi gaussienne, loi de Gauss ou loi de Laplace-Gauss des noms de Laplace (1749-1827) et Gauss (1777-1855), deux mathématiciens, astronomes et physiciens qui l'ont étudiée.

Plus formellement, c'est une loi de probabilités absolument continue qui dépend de deux paramètres : son espérance, un nombre réel noté μ , et son écart type, un nombre réel strictement positif noté σ . La densité de probabilité de la loi normale est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right). \quad (1)$$

La courbe de cette densité est appelée **courbe de Gauss** ou **courbe en cloche**, entre autres. C'est la représentation la plus connue de cette loi. La loi normale d'espérance nulle et d'écart type unitaire est appelée loi normale centrée réduite ou loi normale standard.

Lorsqu'une variable aléatoire X suit la loi normale, elle est dite gaussienne ou normale et il est habituel d'utiliser la notation avec la variance σ^2 . Vous comprenez peut-être maintenant pourquoi nous vous avons obligé à connaître la fonction exponentielle. C'est grâce à cette fonction qu'on modélise cette forme de cloche représentative de la gaussienne. Nous allons essayer de comprendre le pic important suite à notre fonction **density()**.

En statistique, le test de Shapiro-Wilk teste l'hypothèse nulle (aussi appelé hypothèse H_0) selon laquelle un échantillon analysé est issu d'une population normalement distribuée. Nous allons regarder si les ventes sont normalement distribuée pour différents genre

```
> shapiro.test(df$Global_Sales[df$Genre == 'Adventure'])
```

```
Shapiro-Wilk normality test
```

```
data: df$Global_Sales[df$Genre == "Adventure"]
W = 0.30164, p-value < 2.2e-16
```

À vous !

- Interprétez la p value du test de Shapiro sur les jeux d'aventures
- Interprétez la p value du test de Shapiro sur les jeux de stratégie
- Les ventes de *Aventure* et *stratégie* suivent-ils la loi normale ?
- Testez la normalité sur la valeurs *Globale_Sales* de tout les genres de jeux. Que constatez vous ?
- Réalisez un test plus adapté pour tester la distribution normal de *Global_Sales*. Quelles valeurs trouvez vous pour la p-value ?
- Concluez.

1.5 Time to decide

La fonction `plotmeans()` vous permet d'avoir la moyenne et l'intervalle de confiance. L'intervalle de confiance permet d'évaluer la précision de l'estimation d'un paramètre statistique sur un échantillon.

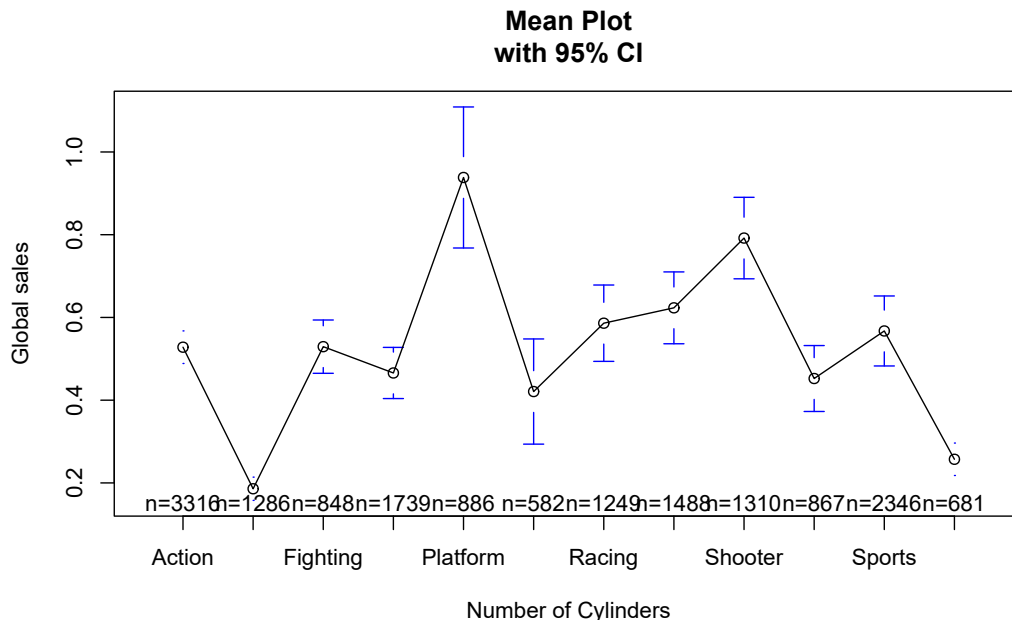


FIGURE 7 – `plotmeans()` sur *Global_Sales*

```
> install.packages(gplots)
> library(gplots)

> plotmeans(df$Global_Sales ~ df$Genre, xlab="Number of Cylinders",
            ylab="Global sales", main="Mean Plot\nwith 95% CI")
Warning messages:
1: In arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd,
:
zero-length arrow is of indeterminate angle and so skipped
2: In arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd,
:
zero-length arrow is of indeterminate angle and so skipped
3: In arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd,
:
zero-length arrow is of indeterminate angle and so skipped
4: In arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd,
:
zero-length arrow is of indeterminate angle and so skipped
5: In arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd,
:
zero-length arrow is of indeterminate angle and so skipped
6: In arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd,
:
```

zero-length arrow is of indeterminate angle and so skipped

À vous !

- a) Citez la formule mathématique permettant de retrouver les valeurs définissant l'intervalles de confiances. Interprétez le graphique 7.
- b) Quel genre n'a pas été bien venu ?
- c) Quel genre à été le mieux venu ?
- d) Quel pouvez vous dire sur le jeux qui à été le mieux vendu ? Pensez vous que ce jeux à vraiment été très populaire ?
- e) Créez un sous ensemble de données, ayant uniquement la liste des jeux depuis 2012.
- f) Affichez la fonction plotmeans des Global_Sales selon les genres.
- g) Quel genre n'a pas été bien venu, depuis 2012 ?
- h) Quel genre à été le mieux venu, depuis 2012 ?
- i) Dans quel genre de jeux pensez-vous qu'il est opportun d'aller ?
- j) Sur quelle plate forme allez-vous proposer votre jeux ?
- k) Dans quel pays allez-vous lancer votre jeux ?
- l) Concluez sur ce jeu de données