Analyse exploratoire (Exploratory Data Analysis)

Mathématiques Générales B
Université de Genève

Sylvain Sardy

28 février 2008

Les données sont des informations quantitatives ou qualitatives.

Ex : $\{1, 0, 0, 1, 1, 1, 1, 0, 1, 0\}$ sont des données de jets d'un dé.

Modélisées comme réalisations d'une v.a., les données peuvent être :

- Discrètes : catégorielles (ex : H/F, P/F) ou ordinales (ex : dé)
- Continues (ex : poids)
- **Univariées** quand on ne mesure qu'un phénomène à la fois.
- Multivariées quand on mesure plusieurs phénomènes conjointement.

	Red	Green	Blue	Orange	Yellow	Brown	Weight
1	15	9	3	NA	9	19	49.79
2	9	17	190	3	3	8	48.98
⋮							

Pour être utiles, les données doivent être :

- vérifiées : données manquantes, aberrantes ?
- **analysées** :
  - résumées avec des chiffres
  - visualisées graphiquement
  - disséquées pour en comprendre la structure et proposer des modèles.
- modélisées : trouver un modèle probabiliste le plus simple possible qui est le plus en adéquation avec la réalité et proche des données.

1. Données univariées

Un casino embauche un statisticien pour trouver de potentiels fraudeurs.

Un jeu consiste à lancer une pièce de monnaie 2 fois et de parier sur le nombre T de Piles.

Des données sont collectées avec :
- Un dé du casino. Un employé est embauché et récolte $N_1 = 1000$ données (2h de travail) : $t_1 = 0, t_2 = 1, t_3 = 2, t_4 = 0, \ldots$
- Un dé amené par un joueur. Observé plus rarement on a $N_2 = 392$ données lors d'une semaine de jeu en 2006 : $t_1 = 0, t_2 = 2, \ldots$

On compte les nombres de fois n_0, n_1, n_2 où $T = 0, 1, 2$.

Comment feriez-vous pour avoir si le dé du joueur ressemble à celui du casino ou s'il est truqué ?

La table des fréquence des données est :

CASINO

T	0	1	2
n_i	231	517	252
$\hat{p}_i = n_i/N_1$	0.231	0.517	0.252

JOUEUR

T	0	1	2
n_i	209	155	28
$\hat{p}_i = n_i/N_2$	0.533	0.395	0.071

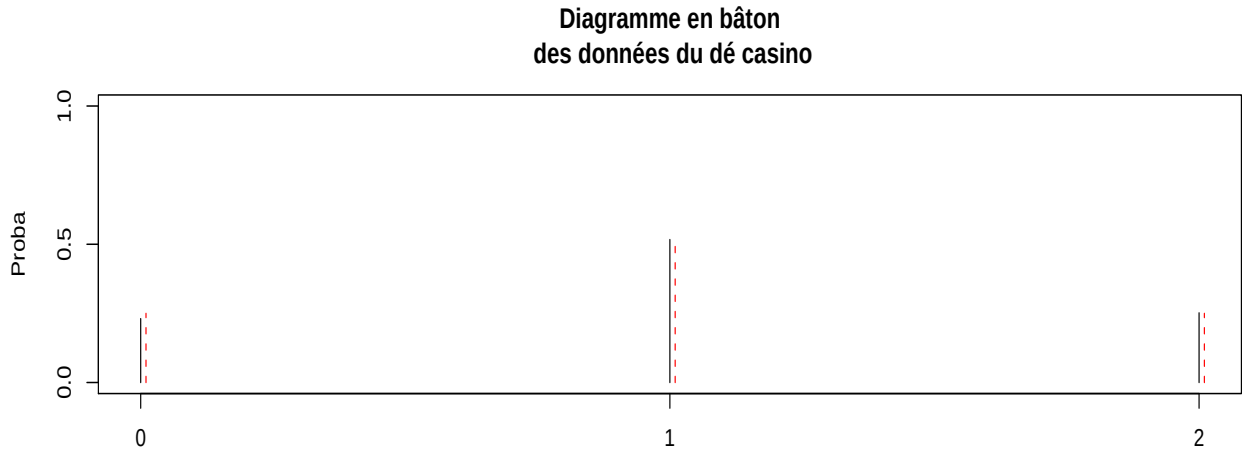


Diagramme en bâton
des données du dé casino
Proba
0.0
0.5
1.0
0
1
2


des données du dé potentiellement truqué
Proba
0.0
0.5
1.0
0
1
2

Application 1 : Le dé du casino est-il équilibré ?

Modélisation probabiliste : soit les variables aléatoires :

- $X_1 \in \{\text{Pile}, \text{Face}\}$ et $X_2 \in \{\text{Pile}, \text{Face}\}$ pour les résultats au premier lancé et au deuxième lancé.
- $T =$ nombre de Pile dans $\{X_1, X_2\}$

On dénote par p la probabilité de Pile :

- Quelles sont les valeurs possibles de $\{X_1, X_2\}$?
- Quelles sont les valeurs possibles de T ?
- Quelles sont les probabilités de réalisation de ces valeurs ?
- Quelles sont les probabilités de réalisation de ces valeurs si le dé est équilibré ?

Le même joueur et le même dé sont observés lors d'un tournoi en 2007, ce qui amène le statisticien à mesurer $N_3 = 114$ lancés de dé.

JOUEUR 2006

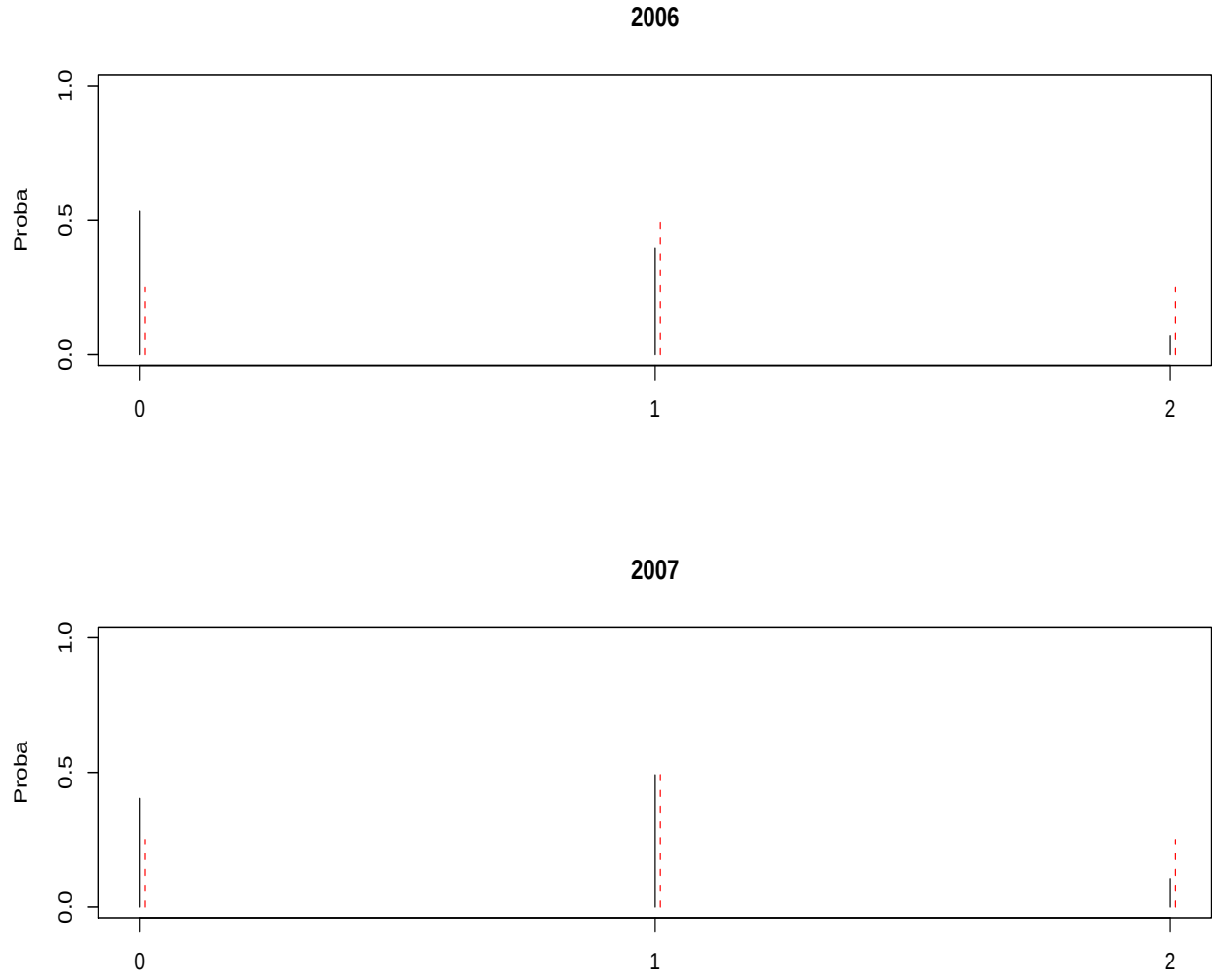
T	0	1	2
n_i	209	155	28
Probabilités estimées $\hat{p}_i$	0.533	0.395	0.071

JOUEUR 2007

T	0	1	2
n_i	46	56	12
Probablités estimées $\hat{p}_i$	0.404	0.491	0.105

DE EQUILIBRE

T	0	1	2
Espérance $\mathrm{E}(n_i)$	$N\frac{1}{4}$	$N\frac{1}{2}$	$N\frac{1}{4}$
Probabilités p_i si $p = \frac{1}{2}$	0.25	0.50	0.25


2006
Proba
0.0
0.5
1.0
0
1
2
2007
Proba
0.0
0.5
1.0
0
1
2

Le rôle des probabilités et des statistiques est de :

- Faire une analyse exploratoire des données pour
- Proposer un modèle probabiliste
- Estimer ce modèle à partir de données
- Vérifier que le modèle colle bien aux données ; sinon, proposer un autre modèle
- Faire de l'inférence, par exemple tester si, pour le dé du joueur, la probabilité d'un Pile est bien $p = 0.5$.

Application 2 : Couleur de M&M's

Le nombre X de M&M's Rouge est mesuré dans $n = 30$ paquets :

15 9 14 15 10 12 6 14 4 9 9 8 12 9 6
4 3 14 5 8 8 9 20 12 8 4 10 5 15 11

Pour les Verts on mesure :

9 17 8 7 3 7 7 11 2 9 11 8 9 7 6
6 5 5 5 9 7 8 2 6 9 6 12 4 11 6

Voyez-vous une différence entre Rouge et Vert ?

Définition : On appelle n la taille de l'échantillon.

L'ensemble fondamental est $\Omega = \{0, 1, 2, \ldots\}$.

Table des fréquence pour les Rouge :

X	0,1,2	3	4	5	6	7	8	9	...
n_i	0	1	3	2	2	0	4	5	
$\hat{p}_i$	0	0.03	0.10	0.07	0.07	0	0.13	0.17	
$\sum_{j \leqslant i} \hat{p}_j$	0	0.03	0.13	0.20	0.27	0.27	0.40	0.57	

X	10	11	12	13	14	15	16,17,18,19	20
n_i	2	1	3	0	3	3	0	1
$\hat{p}_i$	0.07	0.03	0.10	0	0.10	0.10	0	0.03
$\sum_{j \leqslant i} \hat{p}_j$	0.63	0.67	0.77	0.77	0.87	0.97	0.97	1.00

où :

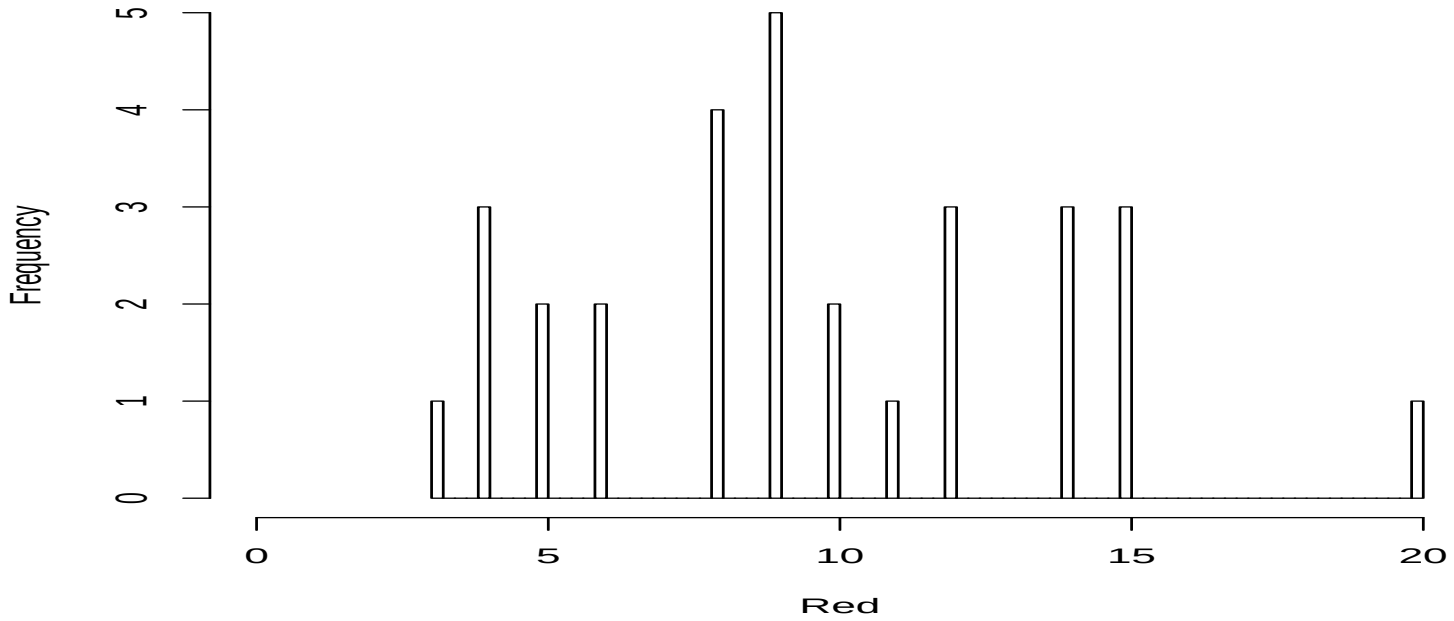
- n_i sont les comptages/fréquences
- $\hat{p}_i = n_i/n$ sont les fréquences relatives/probabilités estimées
- $\sum_{j \leqslant i} \hat{p}_j$ sont les fréquences relatives cumulées.

```
> summary(as.factor(Red))
 3  4  5  6  8  9 10 11 12 14 15 20
 1  3  2  2  4  5  2  1  3  3  3  1
```

```
> round(summary(as.factor(Red))/30,2)
   3    4    5    6    8    9   10   11   12   14   15   20
0.03 0.10 0.07 0.07 0.13 0.17 0.07 0.03 0.10 0.10 0.10 0.03
```

```
> round(cumsum(summary(as.factor(Red))/30),2)
   3    4    5    6    8    9   10   11   12   14   15   20
0.03 0.13 0.20 0.27 0.40 0.57 0.63 0.67 0.77 0.87 0.97 1.00
```

Le diagramme en bâton permet de visualiser la table des fréquences.



Le **mode** est 9 : valeur la plus fréquente.

Données continues

Certaine mesures ne sont par discrètes ou dénombrables.

Exemple : le poids de chaque paquet de M&M's est une variable aléatoire continue. Données arrondies au centième :

49.79 48.98 50.40 49.16 47.61 49.80 50.23 51.68 48.45 46.22 50.43 49.80 46.94
47.98 48.49 48.33 48.72 49.69 48.95 51.71 51.53 50.97 50.01 48.28 48.74 46.72
47.67 47.70 49.40 52.06

L'histogramme est l'équivalent du diagramme à bâton pour les variables/données continues :

- diagramme à bâton $=$ estimateur des probabilités d'une variable aléatoire discrète
- histogramme $=$ estimateur d'une fonction de densité d'une variable aléatoire continue.

(Voir cours "Distributions univariées" au environ de la semaine 10)

Basé sur une partition subjective de l'ensemble fondamental

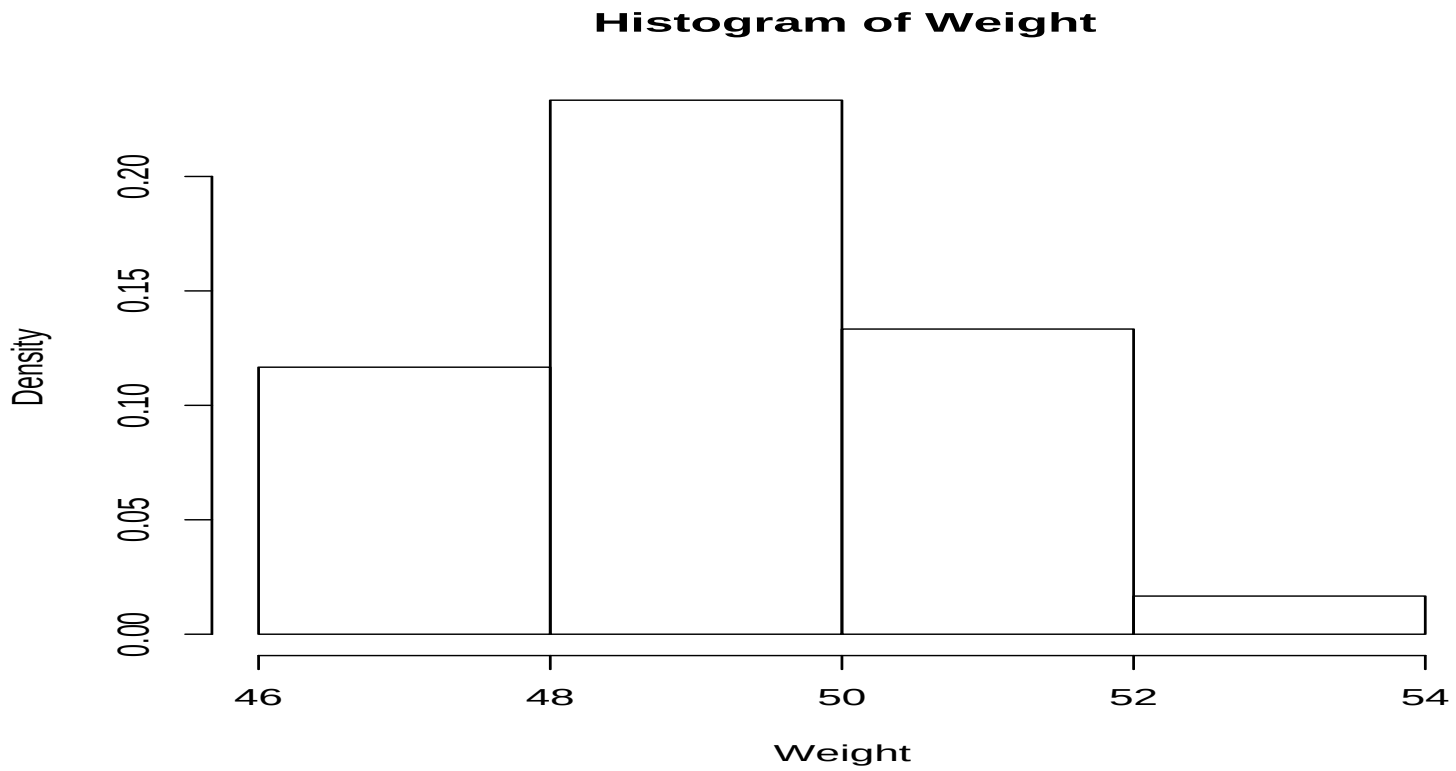
$$\Omega = (0, \infty) = \bigcup_i (b_i, b_{i+1}],$$

l'histogramme est le graphique des **densités** dans chaque intervalle de la partition

	(b_1, b_2)	(b_2, b_3)	(b_3, b_4)	(b_4, b_5)	(b_5, b_6)	(b_6, b_7)
	(0,46)	[46,48)	[48,50)	[50,52)	[52,54)	[54 ,∞)
n_i	0	7	14	8	1	0
$\hat{f}_i$	0	0.12	0.23	0.13	0.02	0

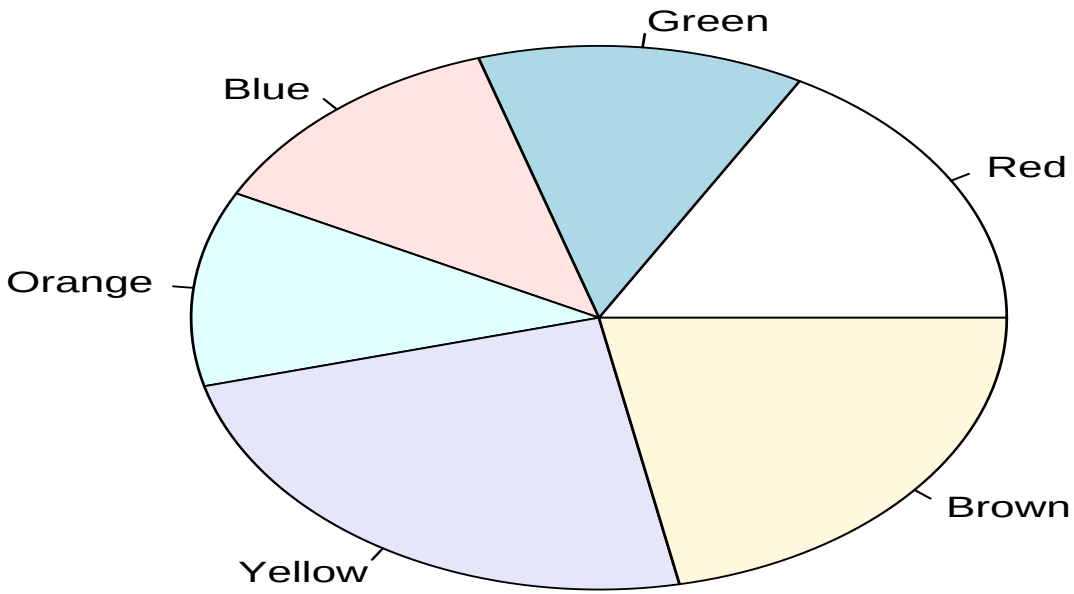
où $\hat{f}_i = \frac{n_i}{n(b_{i+1} - b_i)}$ est la densité estimée.

Illustration de l'histogramme




```
> hist(Weight, breaks=c(46,48,50,52,54), freq=F)
```

Camembert/pie chart



Camembert des couleurs MM's

Statistiques de centralité

Définition : une statistique est une fonction des données $x_1, \ldots, x_n$.

La moyenne : (données discrètes ordinales et continues)

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

Le mode : (données discrètes) réalisation/donnée la plus fréquente (pas forcément unique).

Le mode : (données continues) valeur où la densité a un maximum local (pas forcément unique ; ex. Old Faithfull).

Définition : les statistiques d'ordre $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ sont les données ordonnées, c'est-à-dire

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}.$$

La médiane : (données discrètes ordinales et continues) Valeur telle que 50% des données sont plus petites (et donc que 50% des données sont plus grandes).

$$x_{.5} = \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ est impaire,} \\ \frac{1}{2}(x_{(n/2)} + x_{(1+n/2)}) & \text{si } n \text{ est paire.} \end{cases}$$

Propriété de la médiane : elle est robuste. La robustesse est la propriété d'une statistique à ne pas être influencée de façon trop forte par une 'mauvaise' donnée. C'est à la fois un avantage et un inconvénient.

	Mode	Moyenne	Médiane
Red	9	9.6	9
Green	{6,7,9}	7.4	7
Blue	7	7.2	6.5
Orange	6	6.6	6
Yellow	7	13.8	13.5
Brown	8	12.5	12.5

Pour les irréductibles.

Statistiques de dispersion

Etendue : $x_{(n)} - x_{(1)}$, la différence entre valeurs maximum et minimum.

Définition : le quartile inférieur $x_{.25} = x_{(\lfloor n/4 \rfloor)}$ est la valeur telle que 25% des données sont plus petites. Le quartile supérieur $x_{.75} = x_{(\lceil 3n/4 \rceil)}$ est la valeur telle que 25% des données sont plus grandes.

Etendue interquartile : $\mathrm{EIQ} = x_{.75} - x_{.25}$, différence entre quartiles supérieurs et inférieurs. L'intervalle interquartile $[x_{.25}, x_{.75}]$ contient 50% des données.

Définition : la variance empirique est $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Ecart-type : $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$.

Ecart absolu médian : $\mathrm{mad} = \mathrm{median}(|x - x_{.5}|)$. Robuste ?

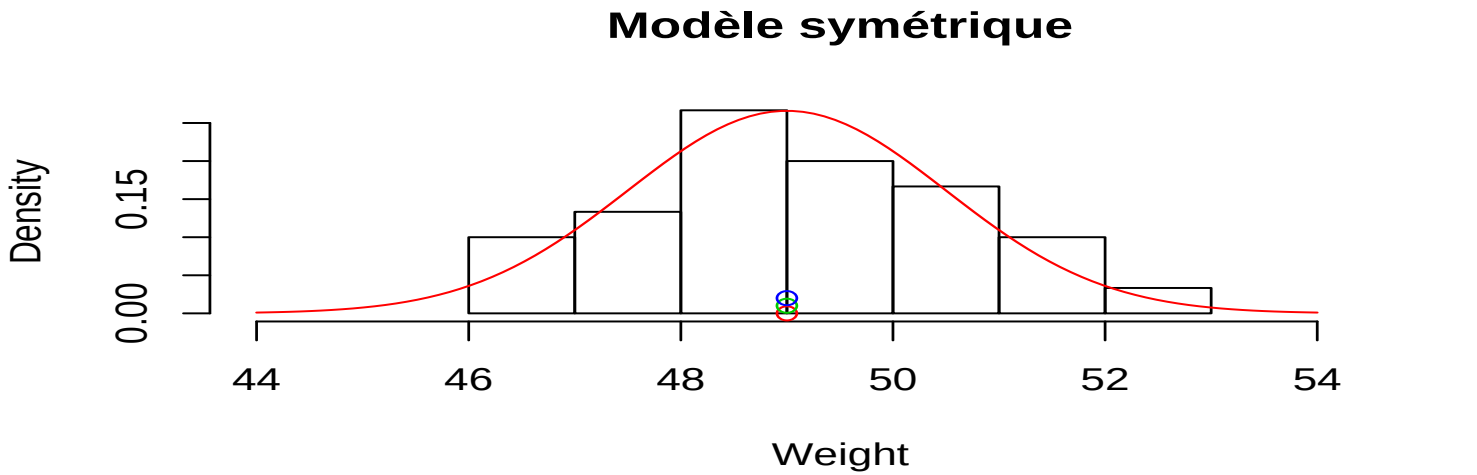
Statistiques d'asymétrie (skewness)

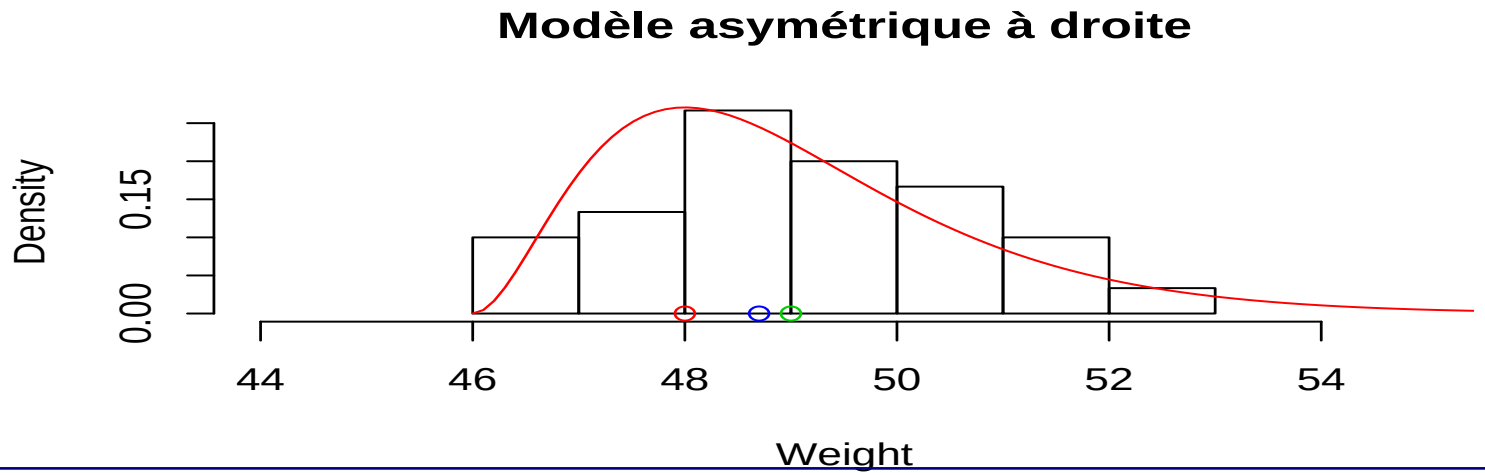
Coefficient d'asymétrie de Fisher : $\hat{\gamma}_1 = \frac{1}{s^3}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}$.

Coefficient d'asymétrie de Yule et Kandall : $u = \frac{(x_{.75} - x_{.5}) - (x_{.5} - x_{.25})}{(x_{.75} - x_{.5}) + (x_{.5} - x_{.25})}$.

La distribution est :
- symétrique quand $\hat{\gamma}_1 = 0$ ou $u = 0$.
- asymétrique à droite quand positive.
- asymétrique à gauche quand négative.

Autre méthode :
- symétrique quand mode=médiane=moyenne.
- asymétrique à droite quand mode<médiane<moyenne.
- asymétrique à gauche quand mode>médiane>moyenne.


Modèle symétrique
Density
0.00
0.15
44
46
48
50
52
54
Weight


Modèle asymétrique à droite
Density
0.00
0.15
44
46
48
50
52
54
Weight

Statistiques d'aplatissement (kurtosis)

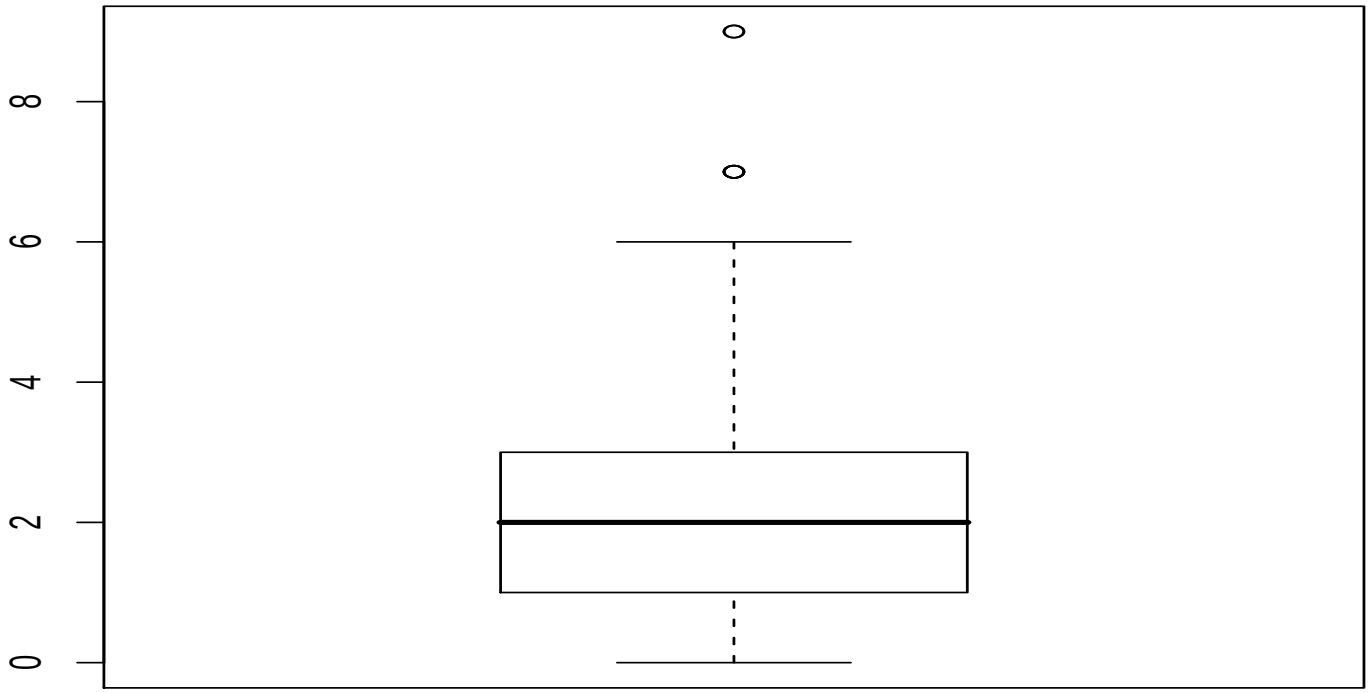
Coefficient d'aplatissement de Fisher : $\hat{\gamma}_2 = \frac{1}{s^4}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{n} - 3$.

La distribution est :
- comme la gaussienne quand $\hat{\gamma}_2 = 0$.
- leptocurtique quand positive.
- platicurtique quand négative.

Les statistiques de moments plus élevés sont peu utilisées car elles ont une grande variance.

Un graphique résume bien ces statistiques : le boxplot.

Le boxplot ou boîte de distribution

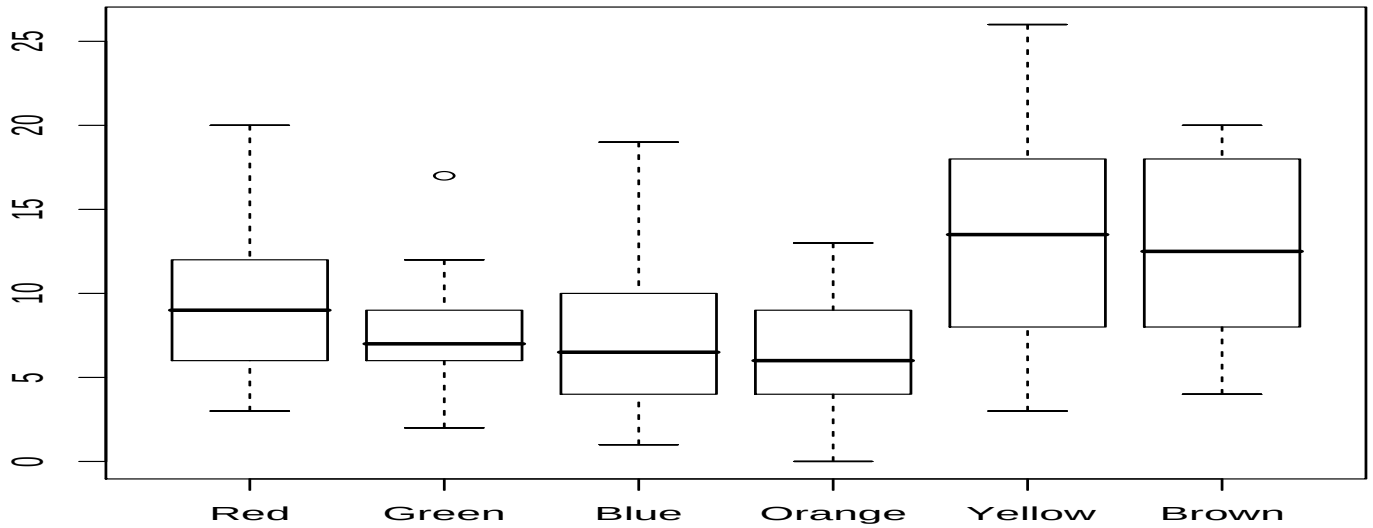


Construction du boxplot :

- la hauteur du rectangle est l'EIQ, le bord bas est à $x_{.25}$ et le bord haut à $x_{.75}$.
- le trait épais au centre du rectangle est la médiane.
- la "moustache" supérieure est la valeur de l'observation la plus proche en deçà de $x_{.75} + 1.5 \times \mathrm{EIQ}$.
- la "moustache" inférieure est la valeur de l'observation la plus proche au delà de $x_{.25} - 1.5 \times \mathrm{EIQ}$.
- les points au delà de ces moustaches sont considérés comme des observations extrêmes, peut-être aberrantes, à regarder de plus prêt.

2. Données multivariées

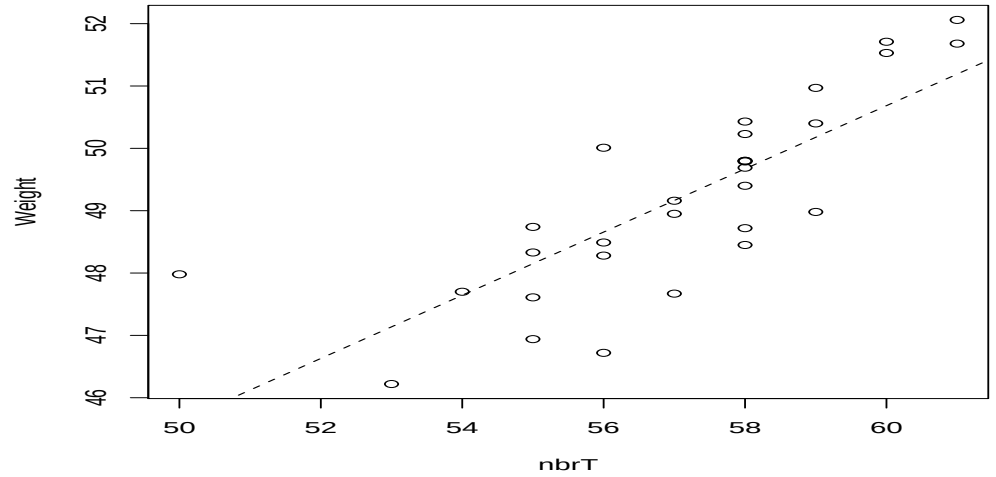
Pour les données M&M's, nous avons analysé les distributions marginales univariées. On peut les comparer :



On doit aussi considérer les distributions conjointes multivariées.

Par exemple : nombres de M&M's x_i dans chaque paquet (toutes couleurs confondues) et poids du paquet y_i pour $i = 1, \ldots, 30$.

Le diagramme de dispersion (ou scatter plot) est le graphe bivarié des couples (x_i, y_i) pour $i = 1, \ldots, n$.



(En pointillé, un modèle de régression linéaire)

Le coefficient de correlation empirique mesure la force de l'association linéaire entre X et Y

$$\hat{\rho}_{X,Y} = \frac{1}{s_X s_Y} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \in [-1, 1].$$

Plus ce chiffre est proche de 1 en valeur absolue, plus l'association linéaire est forte entre les deux variables.

En ingéniérie, on dit que l'association est forte quand $|\hat{\rho}_{X,Y}| \geqslant 0.9$.
En psychologie, on dit que l'association est forte quand $|\hat{\rho}_{X,Y}| \geqslant 0.4$.

Entre `nbrT` et `Weight`, on obtient : $\hat{\rho}_{X,Y} = 0.8$.

Au delà de trois dimensions, il est difficile de visualiser la distribution conjointe des données.

3. Conclusions

L'analyse exploratoire prend du temps. Quand on présente ses résultats.

- Les graphiques doivent rester simples et clairs.
- Tout graphique présenté doit être décrit avec précision : quels sont les axes et les unités, quel est le but du graphique, etc.
- Tout tableau de statistiques doit être décrit avec précision : quels sont les unités, arrondir les statistiques à la décimale reflétant la précision de la statistique.
- Tirer des conclusions de chaque graphique et tableau de statistiques présentés.
- Quand le but est de comparer plusieurs graphiques, garder la même échelle pour tous.