La statistique descriptive
Méthodes de reduction et de représentation des données dans le cas univarié

M. L. Delignette-Muller
VetAgro Sup

5 octobre 2020

VetAgro Sup

Université de Lyon

Objectifs pédagogiques

- Savoir reconnaître le type d'une variable observée.
- Savoir synthétiser et représenter graphiquement des données observées selon le type de la variable.**
- Etre capable d'interpréter les représentations graphiques classiques (dans le cas univarié).
- Savoir juger de la normalité d'une distribution à partir des représentations graphiques classiques.**
- Savoir calculer et interpréter les paramètres statistiques classiques et connaître leurs limites d'utilisation.
- Savoir définir et calculer un intervalle de fluctuation (par ex. pour déterminer des valeurs usuelles).**

** *savoir faire évalué uniquement en S6 après entraînement en TD*

Les types de variables

- Variable **qualitative**
  - **nominale** : modalités **non ordonnées**.
    Ex. : couleur du poil, sexe, ...
  - **ordinale** : modalités **ordonnées**.
    Ex. : évolution de l'état d'un malade (aggravation, état stationnaire, amélioration, guérison), ...
- Variable **quantitative**
  - **discrète** : série **discrète** de nombres.
    Ex. : nombre d'animaux domestiques par foyer, nombre de vétérinaires associés par clinique, ...
  - **continue** : série **continue** de nombres.
    Ex. : poids, durée, taux d'hémoglobine, ...
- Variable **semi-quantitative** (plus compliqué !)
  Ex. : dosage d'un toxique avec une limite de quantification de la méthode analytique, score clinique, ...

Comment bien définir le type d'une variable ?

La bonne question à se poser est :
quelle est la variable observée sur chaque unité d'observation ?
Quelques exemples :

- Etude du poids de chiots à la naissance : unité d'observation = chiot $\Rightarrow$ variable quantitative continue.
- Etude du taux de mortalité des chiots à la naissance dans divers élevages : unité d'observation = élevage $\Rightarrow$ variable quantitative continue.
- Etude du taux de mortalité liée à une pathologie donnée sur un groupe de malades : unité d'observation = individu $\Rightarrow$ variable qualitative nominale (mort / vivant)

Plan

Cas d'une variable qualitative

Etude de la reproduction de chiens de race sur 423 élevages
(données extraites de la thèse vétérinaire de Mathilde Poinssot, Maisons Alfort, 2011)

Une des variables étudiées : le **type de fécondation**

Variable qualitative nominale à trois modalités :

1/ monte naturelle avec un mâle de l'élevage, 2/ monte naturelle avec un autre mâle ou 3/ insémination artificielle.

Données brutes (telles que saisises informatiquement) :

```
    ELEVAGE  FECONDATION
  elevage_1 insemination
  elevage_2   autre_male
  elevage_3 male_elevage
  elevage_4   autre_male
  elevage_5   autre_male
  elevage_6 male_elevage
  elevage_7   autre_male
  elevage_8   autre_male
  elevage_9   autre_male
```

Calcul des effectifs et des fréquences

- Table des **effectifs** n_i pour chacune des classes :

```
autre_male insemination male_elevage
       197          102         124
```

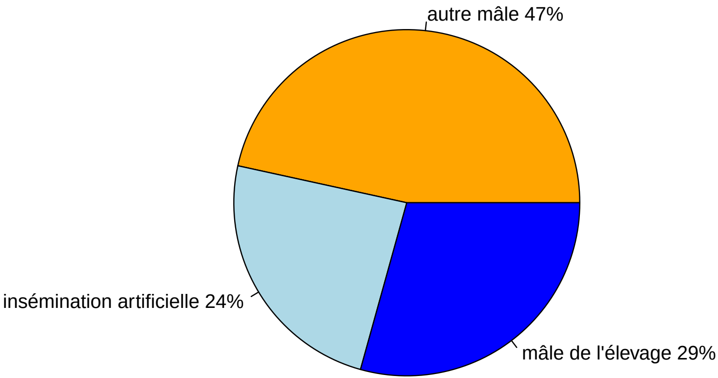
- Table des **fréquences** $f_i = \frac{n_i}{N}$ pour chacune des classes :

```
autre_male insemination male_elevage
     0.466        0.241       0.293
```

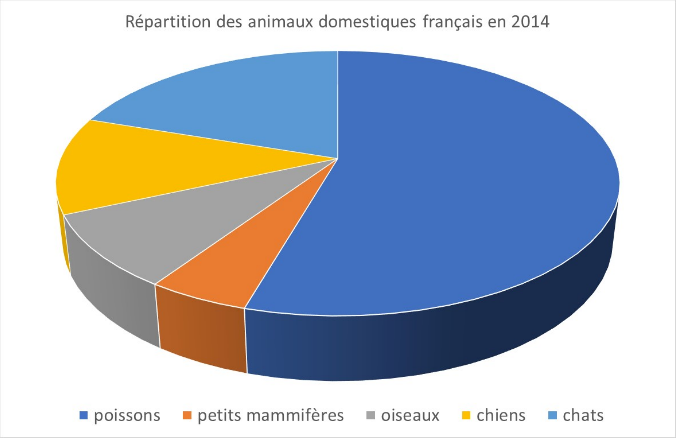
Comment représenter la distribution en fréquences observée ?

Une représentation classique

Le diagramme en secteurs ou camembert.



Représentation manquant parfois de lisibilité



Y a-t-il plus d'oiseaux ou de chiens ? Pas si évident !

Evitez à tout prix les camemberts en relief !

Camembert plus lisible en 2D

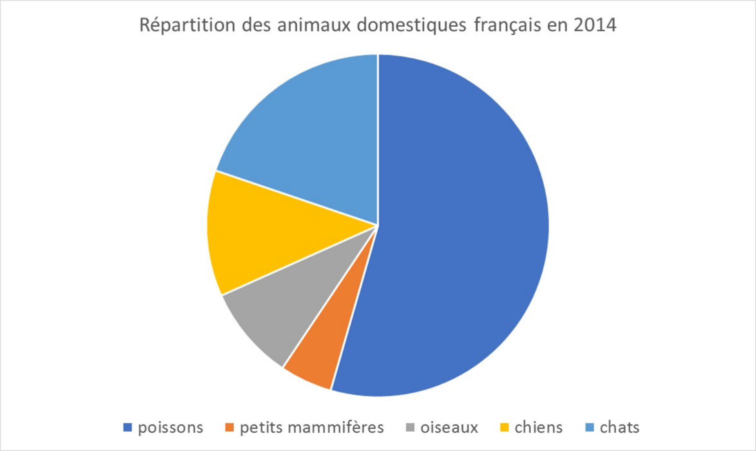
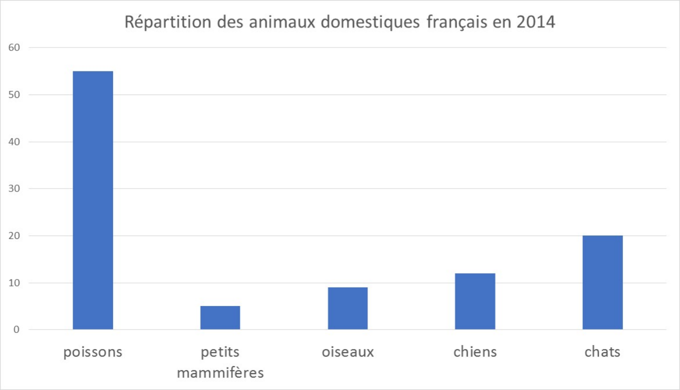
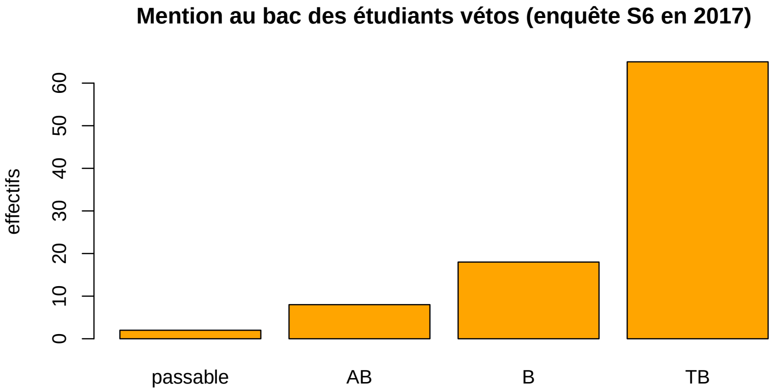


Diagramme en bâtons (en effectifs ou en fréquences) encore plus lisible



Réprésentation préconisée notamment pour les variables qualitatives ordinales.

Diagramme en bâtons à privilégier pour les variables qualitatives ordinales



Cas d'une variable quantitative discrète

Autre variable étudiée dans la thèse précédente sur 998 portées : la **taille de la portée** *i.e.* le **nombre de chiots par portée**

Variable quantitative

avec des **valeurs discrètes** dans l'intervalle $[1, ?]$

Données brutes (telles que saisises informatiquement) :

PORTEE	TAILLE
portee_1	5
portee_2	7
portee_3	3
portee_4	8
portee_5	5
portee_6	6
portee_7	6
portee_8	4
portee_9	6
portee_10	7
portee_11	2

Calcul des effectifs et des fréquences

- Table des **effectifs** n_i pour chacune des classes :

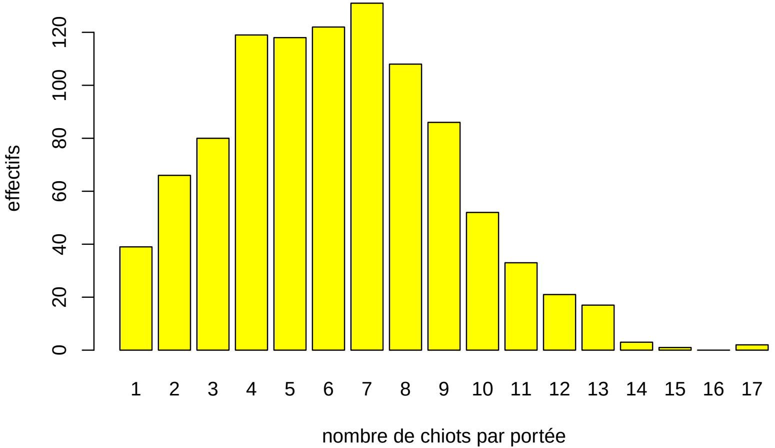
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
39	66	80	119	118	122	131	108	86	52	33	21	17	3	1	0
17															
2															

- Table des **fréquences** $f_i = \frac{n_i}{N}$ pour chacune des classes :

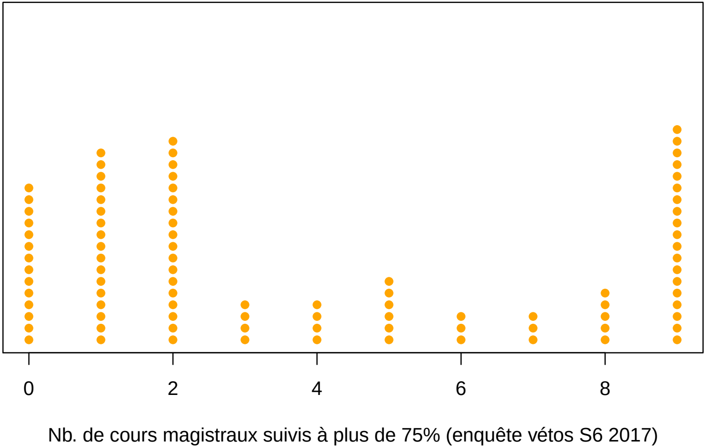
1	2	3	4	5	6	7	8
0.03908	0.06613	0.08016	0.11924	0.11824	0.12224	0.13126	0.10822
9	10	11	12	13	14	15	16
0.08617	0.05210	0.03307	0.02104	0.01703	0.00301	0.00100	0.00000
17							
0.00200							

Cas très similaire à celui d'une variable qualitative ordinale.

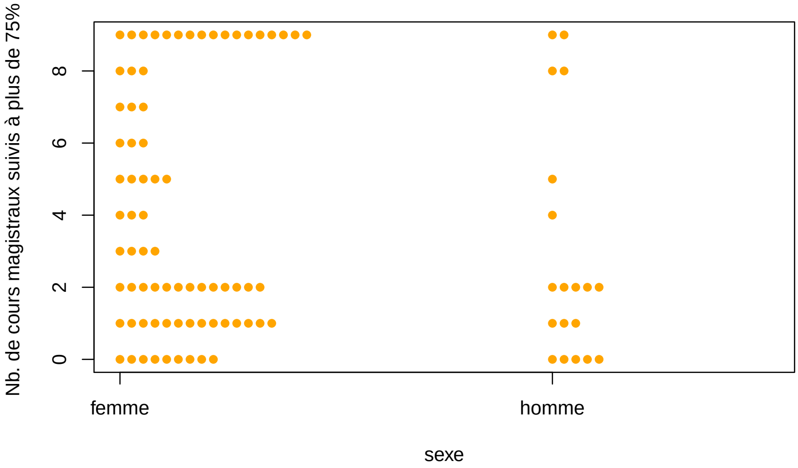
Diagramme en bâtons pour la taille de la portée (en effectifs ou en fréquences)



Graphe des points ("dotplot" ou "stripchart")



Graphe des points ("dotplot" ou "stripchart") en vertical pour comparer plusieurs distributions



Cas d'une variable quantitative continue

Autre variable étudiée dans la thèse précédente sur 928 portées : la **durée de la gestation**

Il s'agit bien d'une variale continue, même si sa mesure est discrète (en jours)

Données brutes (telles que saisises informatiquement) :

PORTEE	DUREE
portee_1	61
portee_2	60
portee_3	62
portee_4	59
portee_5	61
portee_6	62
portee_7	60
portee_8	60
portee_9	62
portee_10	60
portee_11	65

Représentation de la fonction de densité de probabilité

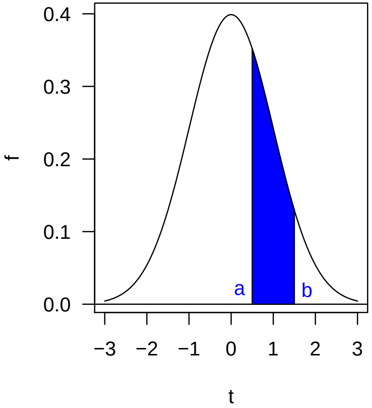
Probabilité d'une valeur donnée $= 0$

Définition d'une **fonction de probabilité** f :

$Pr(a \leq x \leq b) = \int_a^b f(t)dt$

On lit sur le graphe une aire sous la courbe $=$ probabilité d'un intervalle

L'aire globale $= 1$.



Calcul des effectifs et des fréquences par intervalles

- **Définition des intervalles**, par exemple :
  ]45, 50]]50, 55]]55, 60]]60, 65]]65, 70]]70, 75]]75, 80]]80, 85]

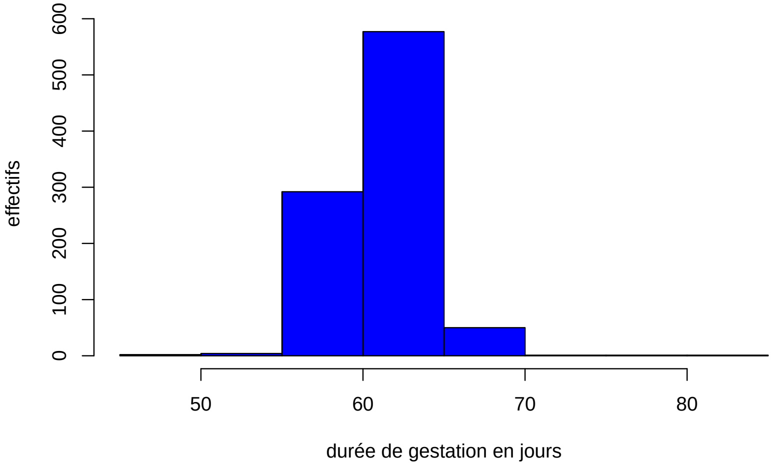
- Calcul des **effectifs** n_i pour chacun des intervalles :

```
]45, 50] ]50, 55] ]55, 60] ]60, 65] ]65, 70] ]70, 75] ]75, 80]
       2        4      292      577       50        1        1
]80, 85]
       1
```

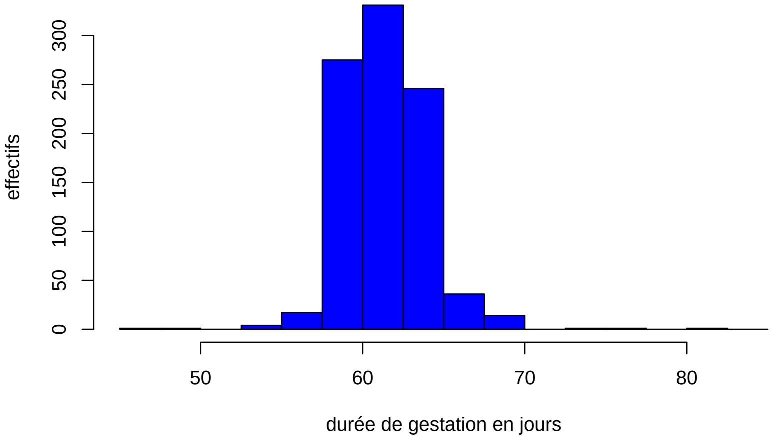
- Calcul des **fréquences** $f_i = \frac{n_i}{N}$ pour chacun des intervalles :

```
]45, 50] ]50, 55] ]55, 60] ]60, 65] ]65, 70] ]70, 75] ]75, 80]
 0.00216  0.00431  0.31466  0.62177  0.05388  0.00108  0.00108
]80, 85]
 0.00108
```

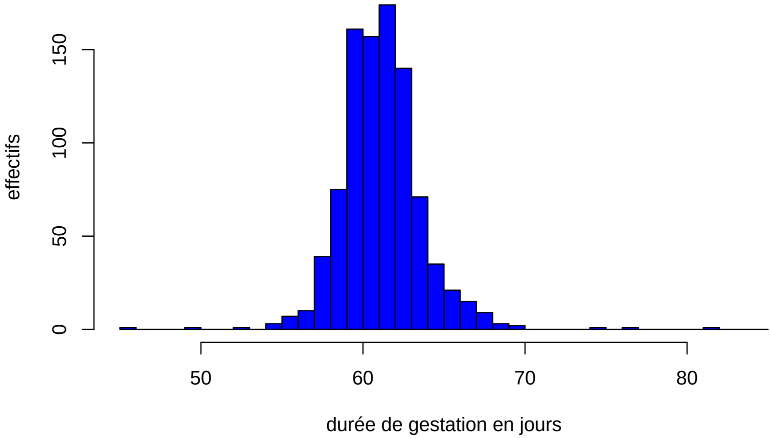
Histogramme de fréquences de la durée de gestation (en effectifs ou en densité de probabilité)



Histogramme de fréquences de la durée de gestation avec des classes plus petites



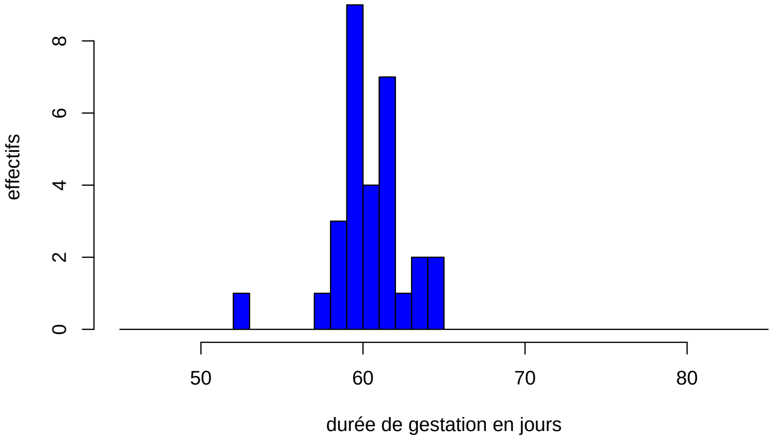
Histogramme de fréquences de la durée de gestation avec des classes encore plus petites



Choix des intervalles

Le choix de la largeur des intervalles dépend beaucoup de l'effectif global.
Plus il est grand, plus on peut se permettre d'affiner l'histogramme en diminuant la largeur des intervalles.
En partant d'un petit effectif l'histogramme devient peu parlant si on prend des intervalles trop étroits.
Avec un très petit effectif, il n'est plus raisonnable de faire un histogramme.
Que donnerait le même histogramme que précédemment si on avait un échantillon de 30 portées ?

Histogramme de la durée de gestation sur 30 portées : intervalles trop étroits !



Histogramme de fréquences de la durée de gestation avec des classes de tailles variables

ATTENTION ! Dans ce cas il faut impérativement le lire en aire sous la courbe et l'axe des y est forcément en densité de probabilité (aire globale $= 1$)

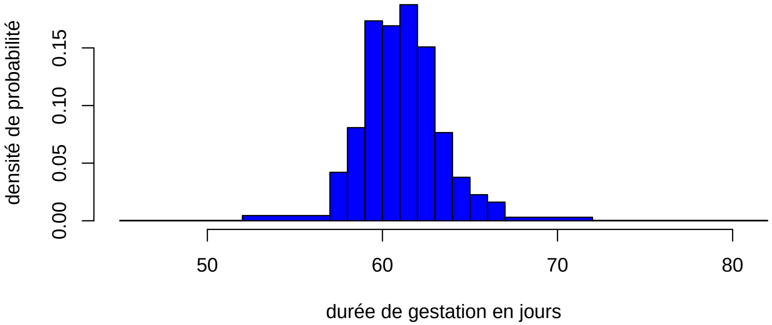
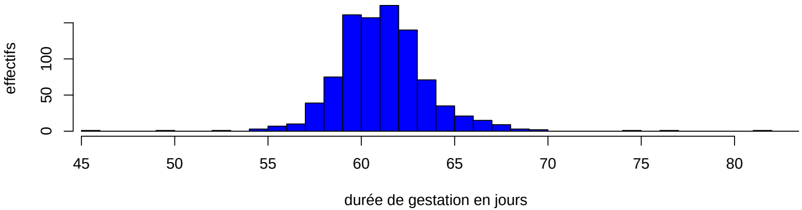
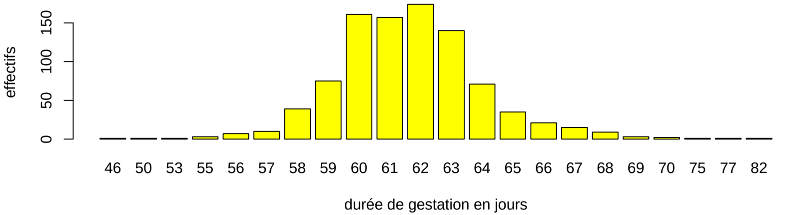


Diagramme en bâton trompeur sur une variale continue

Comparez les deux graphes et trouvez pourquoi le 2^e est inadapté.



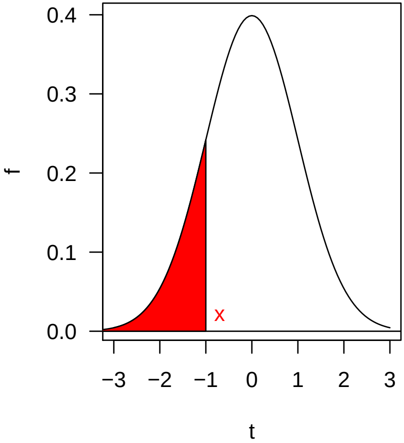


Définition de la fonction de répartition

Définition de la **fonction de répartition** F :

$F(x) = Pr(t \leq x) = \int_{-\infty}^{x} f(t)dt$

Fonction de répartition en x = aire sous la courbe à gauche de x



Représentation de la fonction de répartition

Définition de la **fonction de répartition** F :

$F(x) = Pr(t \leq x) = \int_{-\infty}^{x} f(t)dt$

Fonction de répartition en x = aire sous la courbe à gauche de x

Représentation de la fonction $x \rightarrow F(x)$

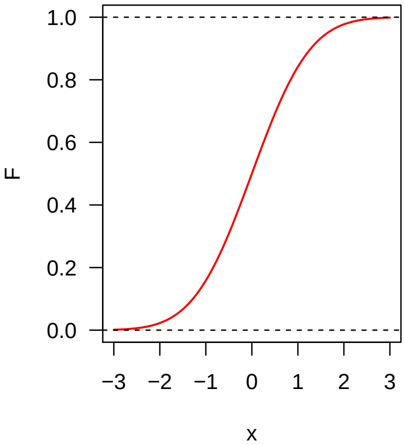


Diagramme des fréquences cumulées

Représentation de la fonction de répartition sur des données observées
Plus de nécessité de définir des classes (intervalles).
On **classe les observations par ordre croissant**,
on attribue **à chaque observation** x_i **son rang** i **dans le classement**,
on peut dire que $F(x_i) = \frac{i}{N}$.
En général on fait une petite correction pour que le graphe parte au-dessus de 0 et arrive en dessous de 1 :
on reporte classiquement les points de coordonnées :
$x = x_i$ **et** $y = \frac{i-0.5}{N}$.

Construction du diagramme des fréquences cumulées de la durée de gestation pour un échantillon de 10 portées

Les valeurs observées ordonnées par ordre croissant

59 59 60 61 61 62 62 63 63 64

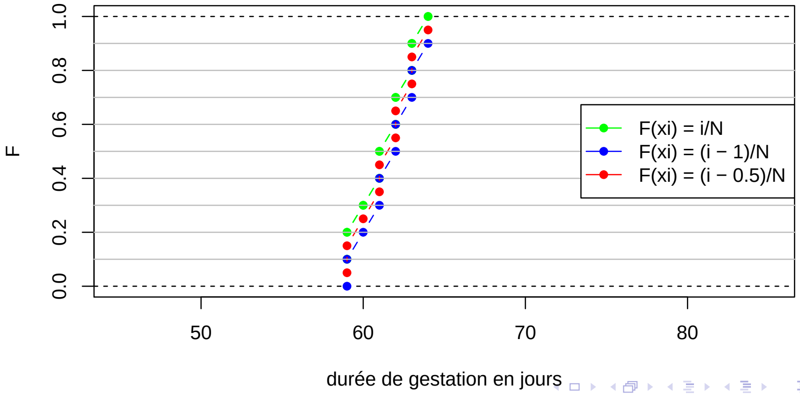


Diagramme des fréquences cumulées pour la durée de gestation (sur les 928 portées)

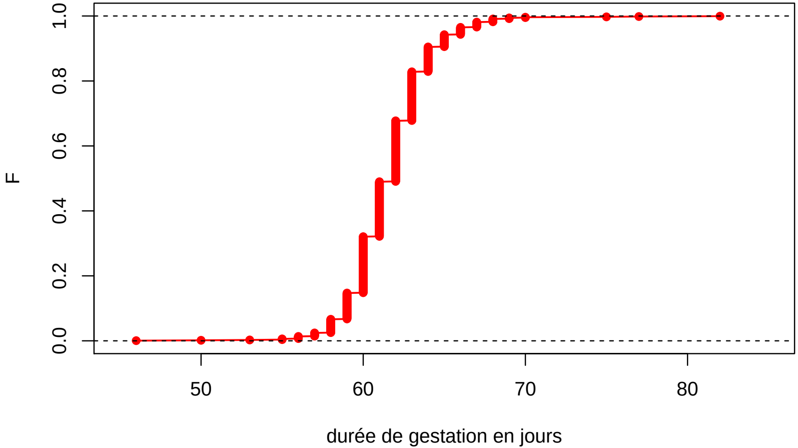


Diagramme en boîte ou boîte à moustaches

Représentation des trois quartiles observés et des valeurs minimale et maximale.

On attribue à chaque observation x_i sa fréquence cumulée comme précédemment

(classiquement $F(x_i) = \frac{i-0.5}{N}$)

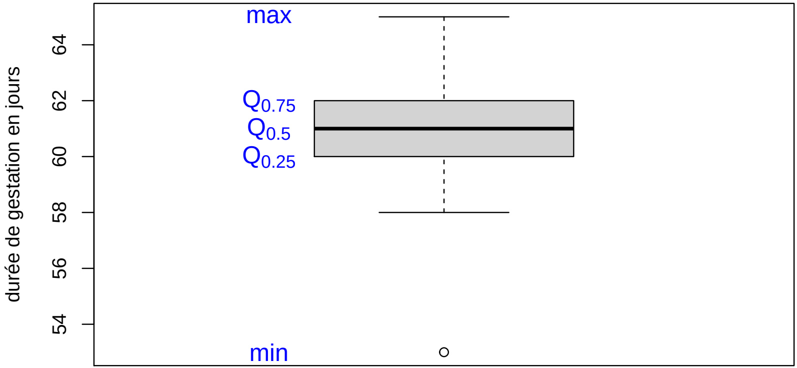
et on définit les valeurs de x correspondant à

$F(x) = 0.25$, 0.5 et 0.75

(diverses méthodes possibles utilisant ou non une interpolation).

- Premier quartile : $F(Q_{0.25}) = 0.25$
- Deuxième quartile (médiane) : $F(Q_{0.5}) = 0.50$
- Troisième quartile : $F(Q_{0.75}) = 0.75$

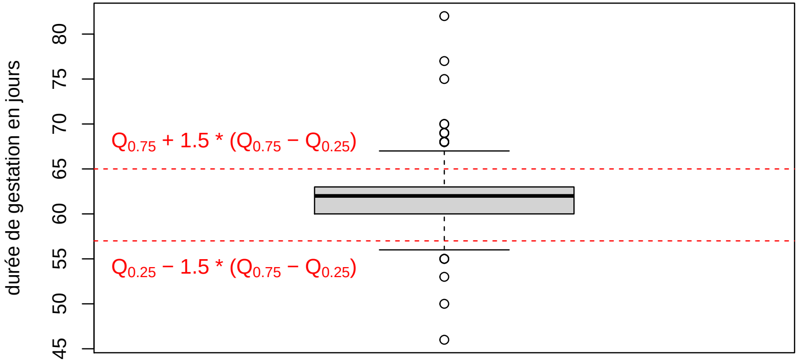
Diagramme en boîte de la durée de gestation sur les 30 portées



Représentation réalisable et parlante même avec peu de données (pas trop peu non plus : pas moins de 7-8 observations)

Diagramme en boîte de la durée de gestation

Il est classique mais non obligatoire de représenter individuellement les valeurs extrêmes (ci-dessous ce que fait le logiciel R par défaut).



Lorsqu'on dispose de vraiment peu d'observations. Exemple de la durée de gestation sur 5 portées.

Le diagramme en boîte n'est pas recommandé. Mieux vaut reporter directement tous les points observés ("dotplot" ou "stripchart").

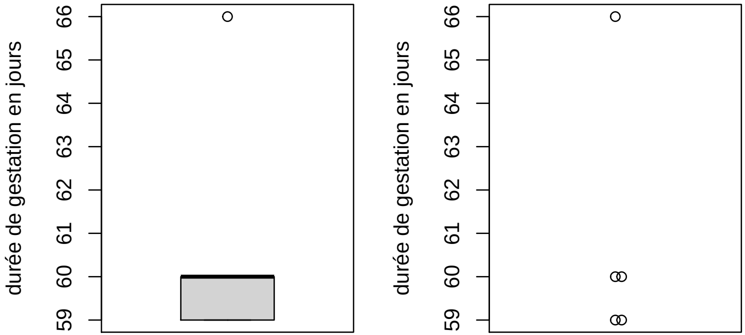


Diagramme Quantile - Quantile ou Q-Q plot

Représentation visant à vérifier la normalité d'une distribution.

On attribue à chaque observation x_i de rang i sa fréquence cumulée :

$$F(x_i) = \frac{i-0.5}{N}.$$

On regarde quelle valeur de u_i dans la loi normale centrée réduite $N(0,1)$ possède la même valeur de F :

$$F_{N(0,1)}(u_i) = F(x_i).$$

Pour chaque observation on reporte un **point d'abscisse** u_i (quantile de la loi normale) et **d'ordonnée** x_i (quantile observé).

Si la loi observée est normale les points sont à peu près alignés.

Construction du Q-Q plot de la durée de gestation pour un échantillon de 10 portées (1)

Les valeurs observées x_i ordonnées par ordre croissant

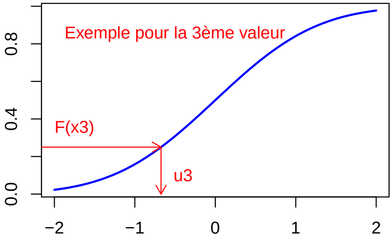
58 60 61 61 62 62 63 64 66 67

Les fréquences cumulées associées $F(x_i)$

0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95

Les valeurs de u_i correspondantes (Loi normale N(0,1))

-1.64 -1.04 -0.674 -0.385 -0.126 0.126 0.385 0.674 1.04 1.64



Construction du Q-Q plot de la durée de gestation pour un échantillon de 10 portées (2)

En ordonnées les valeurs observées x_i ordonnées par ordre croissant

58 60 61 61 62 62 63 64 66 67

En abscisses valeurs de u_i correspondantes

-1.64 -1.04 -0.674 -0.385 -0.126 0.126 0.385 0.674 1.04 1.64

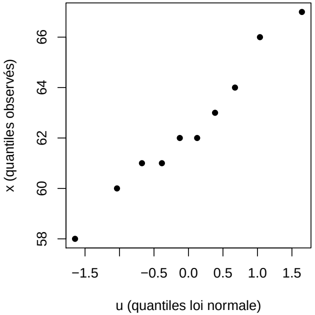
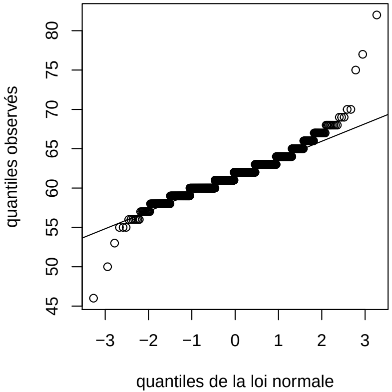


Diagramme Quantile-Quantile de la durée de gestation sur les 928 portées



On voit apparaître un faible écart à la loi normale expliqué partiellement par les valeurs extrêmes.

A RETENIR pour la représentation d'une variable continue

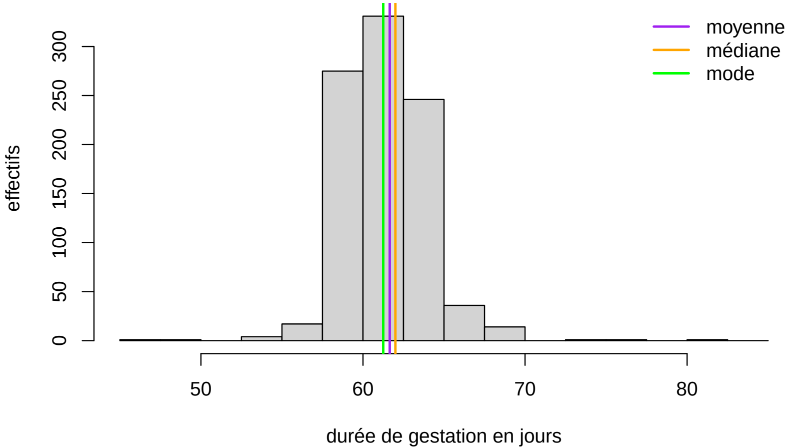
- **Histogramme de fréquences**
  Vision fine de la densité de probabilité - grand nombre de points et définition appropriée de classes nécessaires.
- **Diagramme des fréquences cumulées**
  Visualisation de la fonction de répartition.
- **Diagramme en boîte ("boxplot")**
  Visualisation synthétique de la densité de probabilité - possible même avec un nombre de points modéré (si nombre trop faible, représentation directe des points)
- **Diagramme Quantile-Quantile ("QQ-plot")**
  Vérification de la normalité d'une distribution.

Plan

Paramètres de position
Localisation du centre de la distribution

- **Moyenne**
  sous-entendu moyenne arithmétique classique pouvant être notée de diverses façons :
  $\overline{x} = E(x) = m_x = \frac{1}{N} \sum_{k=1}^{N} x_i$
- **Médiane**
  Deuxième quartile $Q_{0.5}$ (tel que $F(Q_{0.5}) = 0.5$).
  Paramètre robuste (peu sensible aux valeurs extrêmes).
- **Mode**
  Pic de la distribution pouvant être visualisé sur un histogramme comme la valeur centrale de la classe la plus représentée (dépend de la définition des classes).

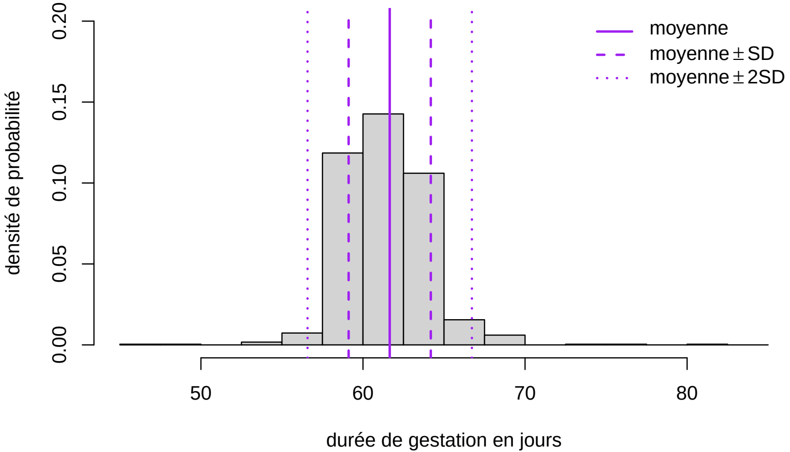
Représentation des paramètres de position sur l'histogramme de la durée de gestation (sur 928 portées)



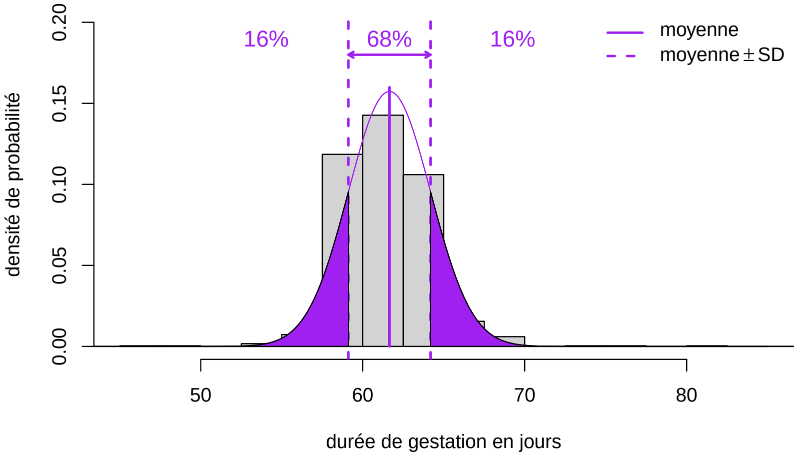
Paramètres de dispersion
Etalement des observations autour de la valeur centrale

- **Variance, écart type, coefficient de variation**
  - **Variance** (moyenne des carrés des écarts à la moyenne) : $V(x) = \frac{1}{N}\sum_{k=1}^{N}(x_i - \overline{x})^2 = E(x^2) - E(x)^2$
  - **Ecart type** (noté souvent SD pour "Standard Deviation") : $SD = \sqrt{V(x)}$
  - **coefficient de variation** : $CV = \frac{SD}{\overline{x}}$
- **Ecart interquartile** (paramètre robuste)
  $EIQ = Q_{0.75} - Q_{0.25}$
  paramètre assez peu utilisé correspondant à la longueur de la boîte dans un diagramme en boîte.

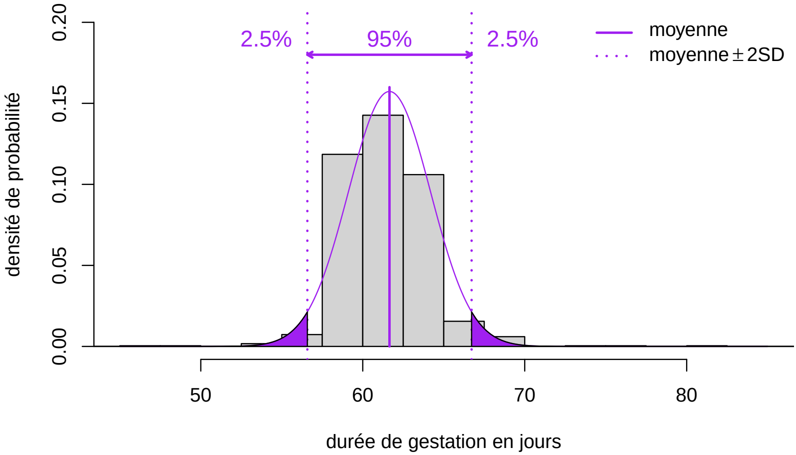
Interprétation de l'écart type sur l'histogramme de la durée de gestation (sur les 928 portées)



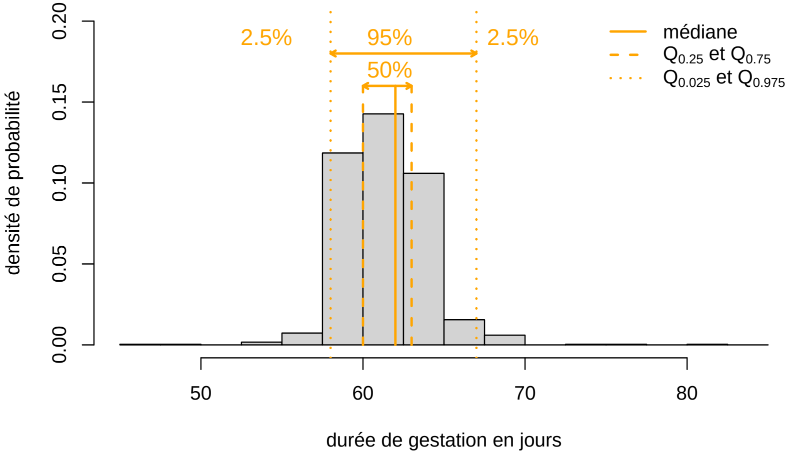
Cas particulier d'une loi observée normale : l'intervalle $\overline{x} \pm SD$ contient 68 % des valeurs



Cas particulier d'une loi observée normale : l'intervalle $\bar{x} \pm 2SD$ contient 95 % des valeurs



Quartiles et des quantiles à 2.5 et 97.5 % sur l'histogramme de la durée de gestation (sur 928 portées)

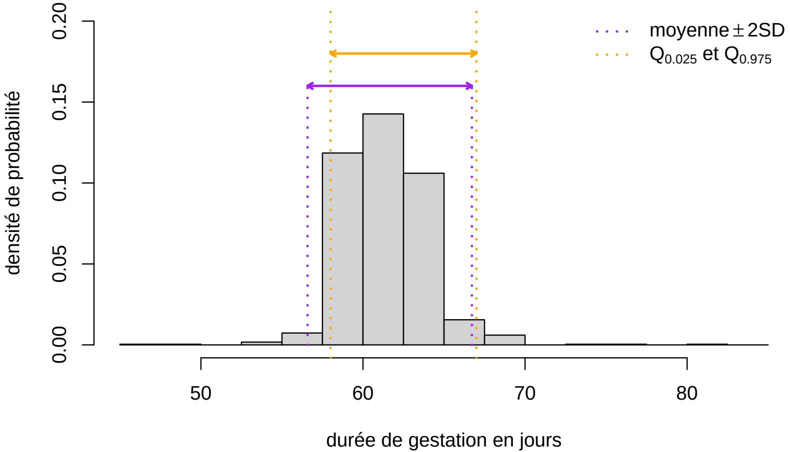


Deux méthodes potentielles pour définir un intervalle de fluctuation

Exemple : valeurs usuelles du taux d'hémoglobine chez le chat ?
Observation du taux d'hémoglobine sur un échantillon de chats sains, puis détermination de l'intervalle contenant 95% des observations (et laissant 2.5% des observations de chacun de ses côtés) = **intervalle de fluctuation à 95%**.

- **Utilisation des quantiles** : $[Q_{0.025}, Q_{0.975}]$
  valable quelle que soit la distribution mais nécessite de nombreuses observations pour une estimation précise.
- **Utilisation de la moyenne et de l'écart type pour une loi normale** : $[\overline{x} - 1.96 \times SD, \overline{x} + 1.96 \times SD]$
  approché souvent par $[\overline{x} - 2 \times SD, \overline{x} + 2 \times SD]$
  ATTENTION, valable uniquement si la loi est proche d'une loi normale.

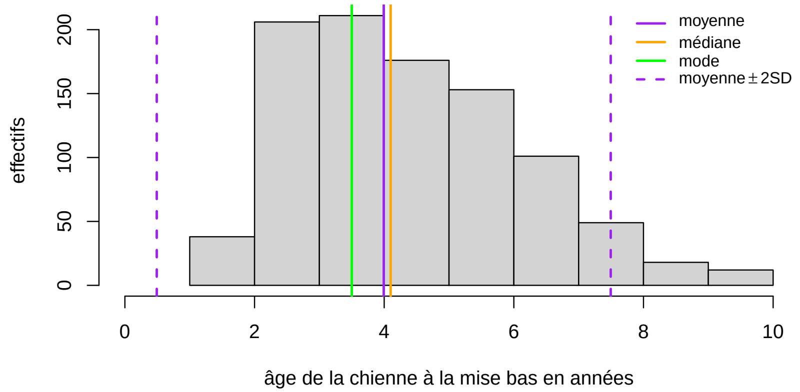
Comparaison des deux méthodes pour la durée de gestation (sur 928 portées)



Limites des paramètres classiques : moyenne et écart type

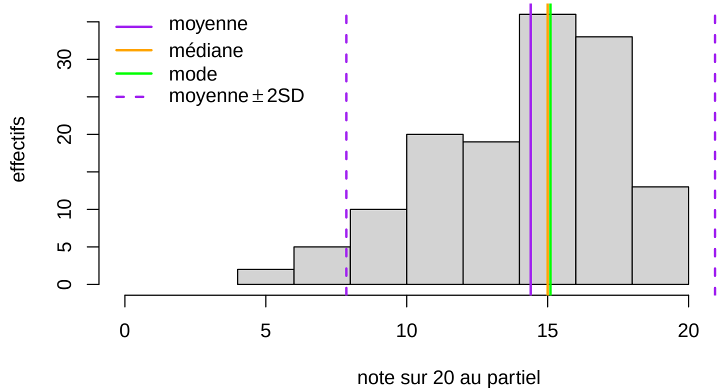
La moyenne et l'écart type résume complètement l'information contenue dans une distribution normale,
mais **ne sont pas appropriés pour résumer une distribution de forme différente**,
d'où l'**importance de représenter graphiquement les données avant tout traitement statistique**.

Histogramme de fréquences de l'âge à la mise bas sur 964 chiennes (même thèse vétérinaire)



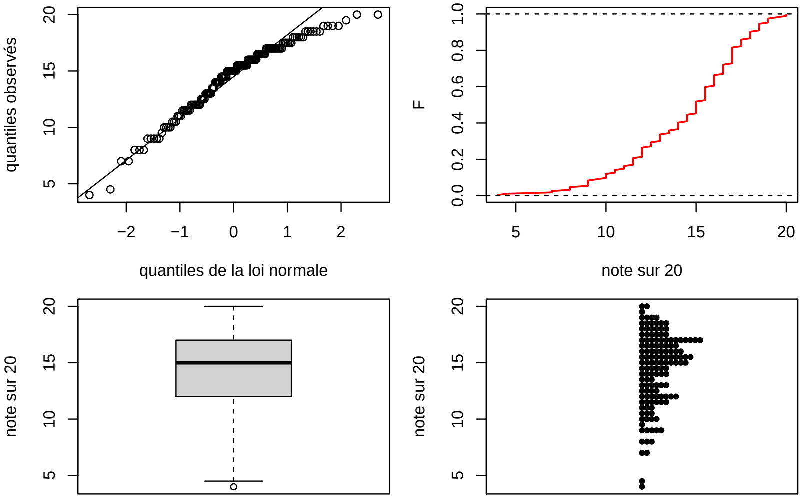
Intervalle $m \pm 2SD$ trompeur : pas de mise bas avant 1 an.

Histogramme de fréquences des notes des étudiants vétérinaires au partiel de biostatistique en juin 2014



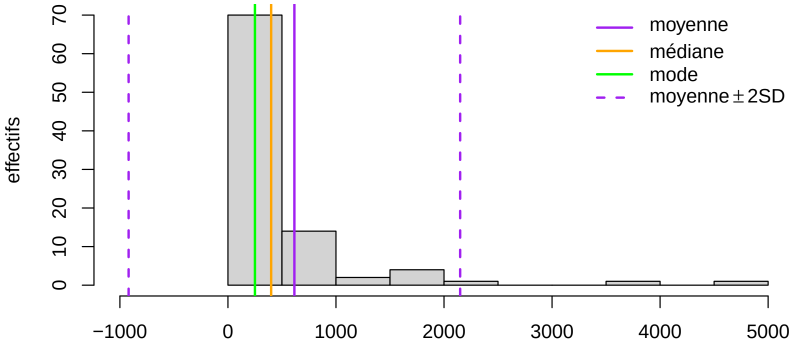
Intervalle $m \pm 2SD$ trompeur : pas de note supérieure à 20.

Autres représentations graphiques de cette distribution



Histogramme de fréquences du coût des soins estimé comme raisonnable pour soigner un chat ou un chien (enquête vétos S6 2017)

Pour de nombreuses distributions les paramètres classiques ont encore moins de sens !



Conclusion

La description des données observées est une étape importante qui doit **IMPERATIVEMENT commencer par une bonne représentation graphique** de la distribution étudiée.
Il convient de **bien réfléchir avant de calculer les paramètres statistiques classiques** (moyenne, variance ou écart type) :
"décrivent-ils bien la distribution observée ?"
Il est parfois plus raisonnable de ne pas résumer les données (très petis effectifs, distributions non normales)