

Module 5 - Statistiques descriptives

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Introduction

Jeu de données statistique

Un jeu de données est habituellement constitué de **plusieurs observations** de **plusieurs variables**.

Exemple d'informations recueillies lors d'un processus de production :

Lot	Luminosité (<i>lux</i>)	Catalyseur	Vitesse du mélangeur	Température (°C)
1	12,0	Oui	Basse	17
2	13,5	Non	Moyenne	15
3	11,5	Non	Haute	14
4	13,0	Oui	Moyenne	18
5	14,0	Non	Basse	21
⋮	⋮	⋮	⋮	⋮

- Variables qualitatives ou quantitatives (traitées différemment en statistique).
- Nous étudierons les variables **quantitatives** dans ce cours.

Décrire avant d'analyser

Une bonne analyse est précédée d'un **résumé** des principales caractéristiques de la distribution des observations.

- mesures de **position**
 - moyenne
 - médiane
 - autres quantiles
- mesures de **dispersion**
 - variance et écart-type
 - écart interquartile
 - étendue
- représentations **graphiques**
 - histogramme
 - diagramme en boîte
 - diagramme quantile-quantile

Description numérique

Tendance centrale : moyenne et médiane

Mesure de position la plus commune : **moyenne échantillonnale**
(version empirique de la moyenne théorique d'une population, μ).

Moyenne échantillonnale

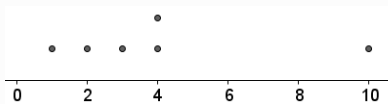
Soit un échantillon observé $x_1, \dots, x_n$.

Sa moyenne échantillonnale est donnée par

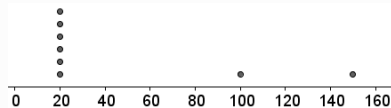
$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Caractéristiques de la moyenne échantillonnale

- Point d'équilibre du diagramme de points
- Fait intervenir toutes les données dans le calcul
- Sensible aux valeurs extrêmes
- Si la somme était répartie également, chaque observation vaudrait $\bar{x}$
- Peut être représentative du groupe... ou non



$$\bar{x} = 4$$



$$\bar{x} = 46,25$$

QUIZ!

La moyenne d'un échantillon de taille $n = 30$ vaut $\bar{x} = 112$.
On retire la plus grande et la plus petite observations.

Vrai ou Faux ?

La moyenne échantillonnale vaut toujours 112.

Médianes théorique et échantillonnale

Autre mesure de tendance centrale : la médiane.

- **Médiane théorique** : associée à la distribution de X , au modèle probabiliste de la population.
- **Médiane échantillonnale** : calculée à partir d'une série d'observations $x_1, \dots, x_n$.

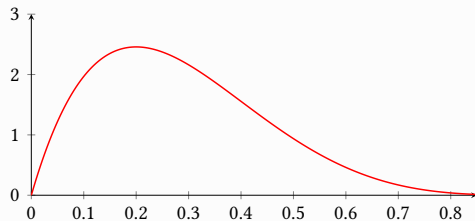
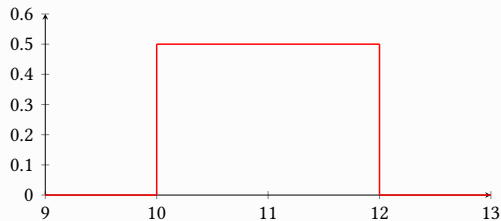
Médiane théorique $\tilde{\mu}$

La médiane de la distribution d'une variable aléatoire X est la valeur $\tilde{\mu}$ telle que

$$P(X \leq \tilde{\mu}) = P(X > \tilde{\mu}) = \frac{1}{2}.$$

QUIZ!

Comment se comparent μ et $\tilde{\mu}$ dans chacune des deux distributions suivantes?



Notation pour l'échantillon ordonné

Pour définir les quantiles échantillonnaires (mesures de position dont la médiane fait partie), on doit introduire les *statistiques d'ordre*.

Statistiques d'ordre

La i -ème statistique d'ordre, notée $x_{(i)}$, est la i -ème observation de l'échantillon *ordonné* (en ordre croissant).

Exemple 1 :	Échantillon dans l'ordre de collecte				Échantillon ordonné			
	x_1	x_2	x_3	x_4	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$
	8	17	2	11				

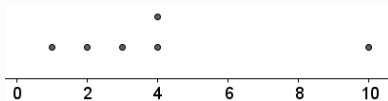
Médiane échantillonnale

Médiane échantillonnale

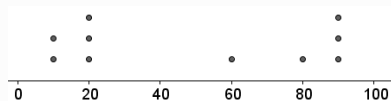
Valeur centrale de l'échantillon *ordonné*, ou la moyenne des deux valeurs centrales dans le cas d'un nombre pair d'observations.

$$\tilde{x} = x_{(\frac{n+1}{2})} = x_{(0,5n+0,5)}$$

- Fait intervenir 1 ou 2 données dans le calcul
- Peu (pas) sensible aux valeurs extrêmes
- Peut être représentative du groupe... ou non



$$\tilde{x} = 3,5$$



$$\tilde{x} = 40$$

QUIZ!

La médiane d'un échantillon de taille $n = 30$ vaut $\tilde{x} = 114$.
On retire la plus grande et la plus petite observations.

Vrai ou Faux ?

La médiane échantillonnale vaut toujours 114.

Quantiles

Quantiles théoriques

Le quantile théorique d'ordre a ($0 \leq a \leq 1$) d'une distribution est la valeur c_a telle que

$$P(X \leq c_a) = a$$

Quantiles échantillonnaires

Il existe plusieurs définitions des quantiles échantillonnaires. Voici celle que nous utiliserons pour le quantile échantillonnal d'ordre a :

$$q_a = x_{(an+0,5)}$$

Si $an + 0,5$ n'est pas entier, on peut faire une interpolation linéaire entre les deux valeurs entières qui l'entourent. On obtient ainsi une moyenne pondérée de $x_{([an+0,5])}$ et $x_{([an+0,5]+1)}$, où $[u]$ est la partie entière de u .

Exemple 2

Considérons un tout petit échantillon :

4 3 8 9 6

- a) Quel est le quantile échantillonnal d'ordre 0,3 de cet échantillon ?

- b) Quel est le quantile échantillonnal d'ordre 0,65 de cet échantillon ?

Quartiles échantillonnaires

Quartiles échantillonnaires

Les 3 quartiles séparent l'échantillon en 4 parties de fréquences égales.

- Le 1^{er} quartile est le quantile d'ordre 0,25 (25^e centile) : $Q_1 = q_{0,25}$.
- Le 2^e quartile est le quantile d'ordre 0,50 (50^e centile) :
 $Q_2 = \tilde{x} = q_{0,50}$ (la médiane !)
- Le 3^e quartile est le quantile d'ordre 0,75 (75^e centile) :
 $Q_3 = q_{0,75}$.

Exemple 3

Soit l'échantillon $x_1, \dots, x_{11}$ suivant : 5, 15, 17, 15, 9, 2, 2, 11, 4, 1, 4.

Calculez ses trois quartiles.

Mesures de dispersion

Deux algorithmes de positionnement par GPS peuvent fournir des localisations moyennes tout aussi précises, mais l'un génère des erreurs beaucoup plus variables d'un point à l'autre. Ainsi, certaines localisations peuvent être très loin de la cible. Ce manque de constance peut être critique dans des applications comme la navigation de drones ou de véhicules autonomes.

En plus des mesures de position, une autre caractéristique importante d'une distribution est **la dispersion des valeurs autour du centre de la distribution**.

Étendue échantillonnale

La mesure la plus simple de la dispersion des données :

Étendue échantillonnale

Différence entre les valeurs maximale et minimale d'un échantillon :

$$R = x_{(n)} - x_{(1)}$$

- Ne tient compte que des 2 valeurs les plus extrêmes
- Pas de lien avec les mesures de tendance centrale

Variance échantillonnale

Rappel : soit X une variable aléatoire quelconque, sa variance est $V(X) = E[(X - \mu)^2]$.

Variance et écart-type échantillonnaux

La **variance échantillonnale** d'un échantillon $x_1, \dots, x_n$ est donnée par

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

L'**écart-type échantillonnal** est $s = \sqrt{s^2}$.

Un truc pour calculer s^2 plus rapidement :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n(\bar{x}^2) \right].$$

Écart interquartile

Une mesure alternative de dispersion est l'écart interquartile.

Écart interquartile

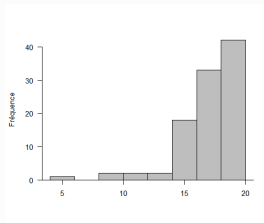
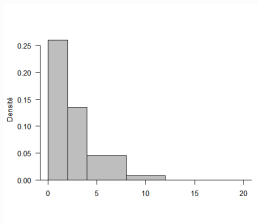
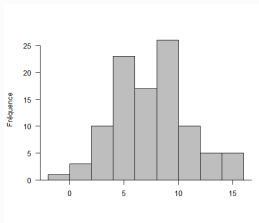
L'écart interquartile d'un échantillon est la distance entre le 1^{er} et le 3^e quartiles :

$$\text{EIQ} = Q_3 - Q_1 = q_{0,75} - q_{0,25}$$

- L'EIQ est l'étendue de la moitié centrale des données.
- Exemple 3 : l'étendue vaut $R = 16$,
l'écart-type échantillonnal est de $s = 5,9$,
l'écart interquartile est $\text{EIQ} = 14 - 2,5 = 11,5$.

Représentation graphique

Histogramme



- L'histogramme sert à **estimer une fonction de densité à partir d'un échantillon**.
- Utile pour déterminer si un certain modèle probabiliste est approprié pour la population d'où provient l'échantillon
- Si l'histogramme est en forme de cloche (**unimodal** et **symétrique**), nous serons à l'aise avec le postulat de normalité de la population.

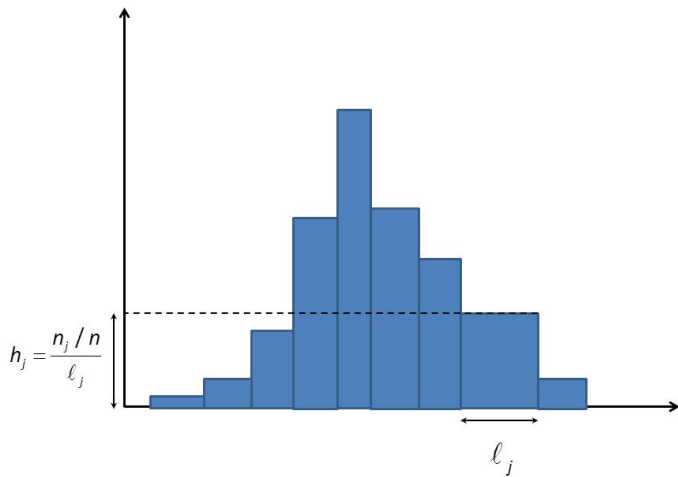
Procédure pour tracer un histogramme

Soit un échantillon $x_1, \dots, x_n$.

- On divise l'ensemble des valeurs possibles de X en environ 5 à 10 classes (intervalles de même largeur ou non).
- Soit K , le nombre de classes. On construit le tableau de fréquences contenant :
 n_j : fréquence (absolue), soit le nombre d'observations qui tombent dans la classe j , et
 $f_j = n_j/n$: fréquence relative de la classe j , pour $j = 1, \dots, K$.
- La surface d'un rectangle doit être proportionnelle à sa fréquence relative. Soit ℓ_j , la largeur de la classe j , alors $\ell_j \cdot h_j = c \cdot f_j$, avec c constant.
- Classes égales : hauteurs égales aux fréquences absolues ou relatives.
- Classes inégales : histogramme de densité ($c = 1$), où

$$h_j = f_j/\ell_j$$

Hauteur des rectangles d'un histogramme



Exemple 4

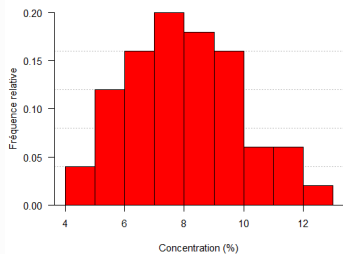
Cinquante mesures de la concentration de phosphates (en % du poids total) sont prises lors de recherches sur de nouvelles façons de produire des détergents. On obtient les observations suivantes :

4,0	4,6	5,4	5,4	5,7	5,7	5,8	5,9	6,5	6,6
6,6	6,7	6,8	6,8	7,0	7,0	7,1	7,1	7,2	7,3
7,4	7,5	7,5	7,5	7,8	7,8	8,1	8,2	8,2	8,3
8,7	8,8	8,8	8,9	8,9	9,1	9,3	9,4	9,5	9,6
9,6	9,7	9,7	10,4	10,4	10,5	11,1	11,4	11,5	12,3

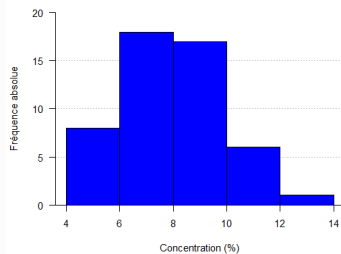
Tracer un histogramme dont les classes sont de largeur égale, et un autre dont les classes ont des longueurs variables.

Histogramme dont les classes sont de largeur égale

Avec des classes de même largeur de la forme $(x_{min}, x_{max}]$, l'histogramme aura l'allure suivante :



ou



Exemple 4 (suite)

Complétons le tableau de fréquences qui définit 8 classes de valeurs.

Classe (j)	Largeur (ℓ_j)	Fréq. (n_j)	Fréq. rel. (f_j)	Hauteur (h_j)
(4, 0; 6, 0]				
(6, 0; 7, 0]		8		
(7, 0; 8, 0]		10		
(8, 0; 8, 5]	0,5	4	0,08	0,16
(8, 5; 9, 0]	0,5	5	0,10	0,20
(9, 0; 10, 0]	1	8	0,16	0,16
(10, 0; 11, 0]	1	3	0,06	0,06
(11, 0; 13, 0]	2	4	0,08	0,04

Histogramme avec classes de largeur inégale

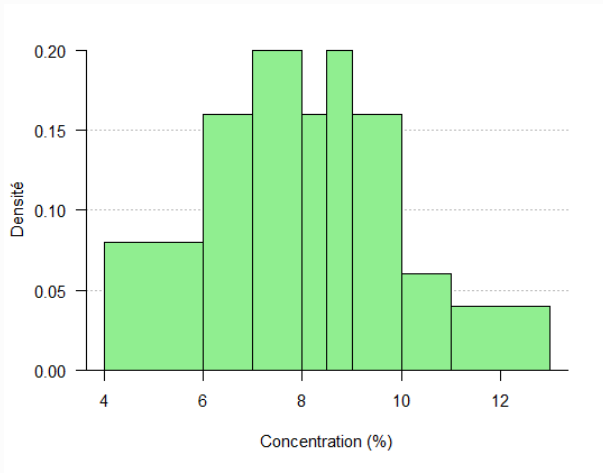
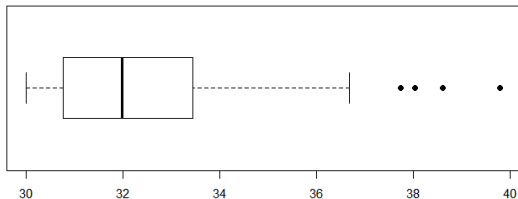


Diagramme en boîte

Le **diagramme en boîte** (*boxplot* ou diagramme de quartiles) permet aussi d'illustrer la distribution d'une série d'observations.

Un diagramme en boîte a l'allure suivante :



Procédure de construction d'un diag. en boîte

Un diagramme en boîte peut être tracé à la verticale ou à l'horizontale. Il est constitué :

- d'un axe gradué
- d'une boîte aux extrémités en $Q_1 = q_{0,25}$ et $Q_3 = q_{0,75}$
- d'une ligne centrale au niveau de la médiane Q_2
- on calcule (**on ne trace PAS**) des bornes théoriques inférieure et supérieure :

$$b_{inf} = q_{0,25} - 1,5EIQ \quad (1)$$

$$b_{sup} = q_{0,75} + 1,5EIQ \quad (2)$$

- d'un trait ("moustache") de Q_3 jusqu'à **la plus grande valeur de l'échantillon inférieure à b_{sup}**
- d'un trait ("moustache") de Q_1 jusqu'à **la plus petite valeur de l'échantillon supérieure à b_{inf}**

Procédure (suite)

- de points pour chaque observation à l'extérieur de l'intervalle $[b_{inf}, b_{sup}]$

Les observations représentées par des points sur un boxplot sont parfois appelées *données aberrantes* ou *données extrêmes*, puisqu'elles sont très peu plausibles sous une loi normale.

Exercice : Remplacez les quartiles échantillonnaires par les quartiles théoriques sous une loi $N(\mu, \sigma^2)$ dans les équations (1)-(2) et calculez $P(X < b_{inf} \text{ ou } X > b_{sup})$ si $X \sim N(\mu, \sigma^2)$. Quelle est la probabilité d'observer une valeur extrême, i.e. hors de $[b_{inf}, b_{sup}]$?

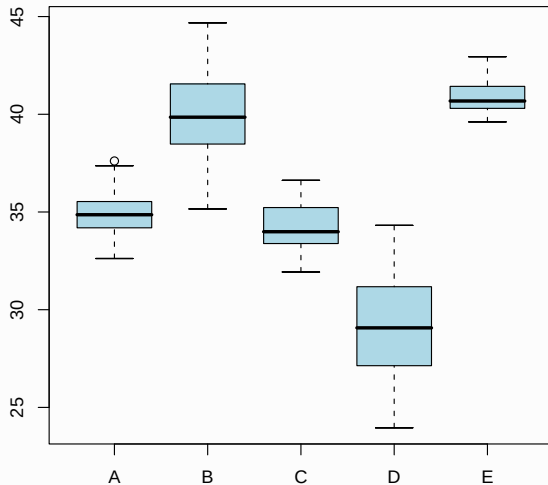
Utilité du diagramme en boîte

Le diagramme en boîte permet d'évaluer en un coup d'œil certaines propriétés de la distribution des observations :

- le centre de la distribution (mesuré par la médiane);
- la dispersion (mesurée par l'écart inter-quartile);
- la symétrie (l'écart entre le 1er quartile et la médiane est-il le même que l'écart entre la médiane et le 3e quartile ? les moustaches sont-elles de même longueur ?);
- si certaines observations sont peu plausibles sous la loi normale (trop extrêmes);

Mais surtout idéal pour comparer plusieurs échantillons simultanément (voir page suivante).

Diagrammes en boîte juxtaposés



Exercice (retour à l'exemple 4)

Construire à la main le diagramme en boîte des concentrations de phosphates, et montrer qu'il a l'allure suivante :

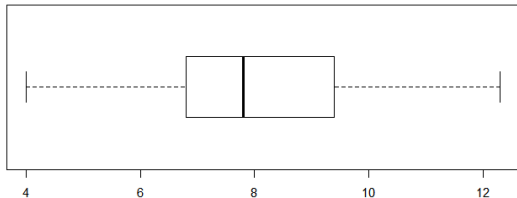


Diagramme quantile-quantile normal

- But : vérifier si la loi normale est un bon modèle pour la population à l'étude
- Principe : les quantiles de l'échantillon sont comparés aux quantiles auxquels on s'attendrait sous une loi normale.
- Technique (grosso modo) : on associe à chaque statistique d'ordre $x_{(i)}$ la valeur du quantile théorique normal d'ordre $i/(n + 1)$. On trace les points sur un graphique, et les points seront alignés si le modèle théorique est bon.

Exemple 4 des concentrations de phosphate

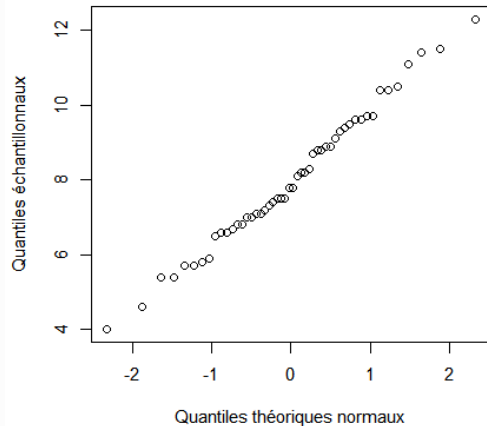
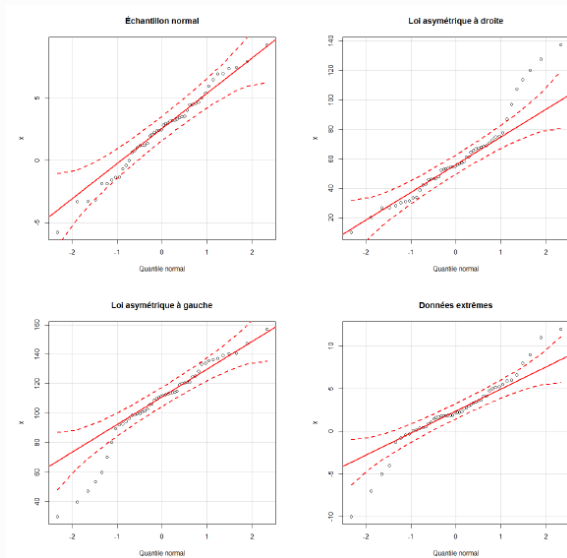


Diagramme quantile-quantile normal



Réponses aux exemples

1. $\{x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}\} = \{2, 8, 11, 17\}$

2. $q_{0,3} = 4$

$$q_{0,65} = 7,5$$

3. $Q_1 = 2,5, Q_2 = 5, Q_3 = 14$

4.

Classe (j)	Largeur (ℓ_j)	Fréq. (n_j)	Fréq. rel. (f_j)	Hauteur (h_j)
(4, 0; 6, 0]	2	8	0,16	0,08
(6, 0; 7, 0]	1	8	0,16	0,16
(7, 0; 8, 0]	1	10	0,20	0,20