MTH2302D: Statistique descriptive

synthèse et présentation de données

Wissem Maazoun

École Polytechnique de Montréal,
Département de Mathématiques et de génie industriel

Automne 2011

Introduction

- La statistique fait intervenir la collecte, la présentation et l'analyse de données, ainsi que leur utilisation dans le but de résoudre des problèmes.
- D'une autre manière, la statistique est une discipline scientifique dont le but est :
  - de planifier et recueillir des données pertinentes ;
  - d'extraire l'information contenue dans un ensemble de données ;
  - de fournir une analyse et interprétation des données afin de pouvoir prendre des décisions
- La statistique utilise :
  - des notions de probabilités ;
  - des notions de mathématiques.

Terminologie statistique

L'univers est l'ensemble d'êtres ou d'objets sur lesquels porte l'étude statistique. Un univers est généralement infini.

La variable est la caractéristique (ou le critère) suivant laquelle l'univers est étudié. Une variable peut être qualitative ou quantitative.

La population est l'ensemble de toutes les mesures de la variable dans l'univers considéré. Une population (tout comme un univers) est généralement infinie.

L'unité expérimentale est un être ou un objet de l'univers. C'est l'élément auprès duquel la variable est mesurée.

Wissem Maazoun | MTH2302D: Statistique descriptive | 3 / 37

Terminologie statistique (suite)

Un échantillon est un sous-ensemble d'unités expérimentales prises au hasard dans l'univers.
On distingue
- un échantillon d'unités expérimentales ;
- un échantillon de la population.

Un paramètre est une mesure caractérisant la variable dans la population. Les paramètres sont généralement inconnus.

Une statistique est une mesure caractérisant la variable dans un échantillon.

Définition

Ce chapitre traite de la statistique descriptive. Il s'agit d'un ensemble de méthodes (représentations graphiques et calculs de caractéristiques numériques) permettant de faire une synthèse statistique de données. Les données à examiner proviennent généralement d'un échantillon.

Représentation graphique

Introduction

Il existe plusieurs méthodes de représentations graphiques et tabulaires de données. Nous représentons ici un certain nombre de ces graphiques ainsi que le but essentiel recherché par chacun. Les diagrammes servent à illustrer une série de données, mais certains servent pour deux séries.

Le diagramme "tige-feuille" ("Stem-and-leaf")

Il s'agit d'une représentation des données dans laquelle chaque observation est divisée en deux parties :
- une tige (stem) ;
- une feuile (leaf).

Pour l'ensemble des observations, on obtient des lignes horizontales (ou verticales), chaque ligne débutant par une tige.

Représentation graphique

Exemple

On dispose des données suivantes :

223 241 245 265 268 267 228 301 300 301 321 282 286 288.

Dresser un diagramme tige-feuille pour ces données.

La distribution des fréquences

Il s'agit d'un type de représentation tabulaire dans laquelle les données sont regroupées par classes.

- Les limites des classes doivent-être bien définies de sorte que chaque observation appartienne à une et une seule classe.
- Le nombre de classe ne peut être ni trop bas ni trop élevé (entre 6 et 12 classes environ).

Représentation graphique

Exemple

On dispose des données suivantes sur l'indice d'octane de 80 spécimens de carburant :

88, 5	94, 7	88, 2	88, 5	93, 3	87, 4	91, 1	90, 5
87, 7	91, 1	90, 8	90, 1	91, 8	88, 4	92, 6	93, 7
83, 4	91, 0	88, 3	89, 2	92, 3	88, 9	89, 8	92, 7
86, 7	94, 2	98, 8	88, 3	90, 4	91, 2	90, 6	92, 2
87, 5	87, 8	94, 2	85, 3	90, 1	89, 3	91, 1	92, 2
91, 5	89, 9	92, 7	87, 9	93, 0	94, 4	90, 4	91, 2
88, 6	88, 3	93, 2	88, 6	88, 7	92, 7	89, 3	91, 0
100, 3	87, 6	91, 0	90, 9	89, 9	91, 8	89, 7	92, 2
95, 6	84, 3	90, 3	89, 0	89, 8	91, 6	90, 3	90, 0
93, 3	86, 7	93, 4	96, 1	89, 6	90, 4	91, 6	90, 7

Représentation graphique

	Frequency table: IND_OCT (octane) Count	Cumulative Count	Percent	Cumulative Percent
83,00 <= X < 85,00	2	2	2,50000	2,5000
85,00 <= X < 87,00	3	5	3,75000	6,2500
87,00 <= X < 89,00	17	22	21,25000	27,5000
89,00 <= X < 91,00	23	45	28,75000	56,2500
91,00 <= X < 93,00	21	66	26,25000	82,5000
93,00 <= X < 95,00	10	76	12,50000	95,0000
95,00 <= X < 97,00	2	78	2,50000	97,5000
97,00 <= X < 99,00	1	79	1,25000	98,7500
99,00 <= X < 101,0	1	80	1,25000	100,0000

Figure: La distribution des fréquences obtenue par Statistica

Wissem Maazoun MTH2302D: Statistique descriptive 9 / 37

Représentation graphique

L'histogramme

On peut illustrer la distribution des fréquences d'une variable à l'aide d'un histogramme. Il s'agit d'un graphique où chaque classe est représentée par un rectangle dont la surface (aire) est proportionnelle à la fréquence relative de la classe. Deux posibilités :

- les fréquences relatives sont utilisées, on obtient un histogramme. La courbe formée de segment de droites joignant les milieus des sommets des rectangles est appelée le polygone des fréquences ;
- les fréquences cumulées sont utilisées, on obtient l'histogramme des fréquences cumulées.

Représentation graphique

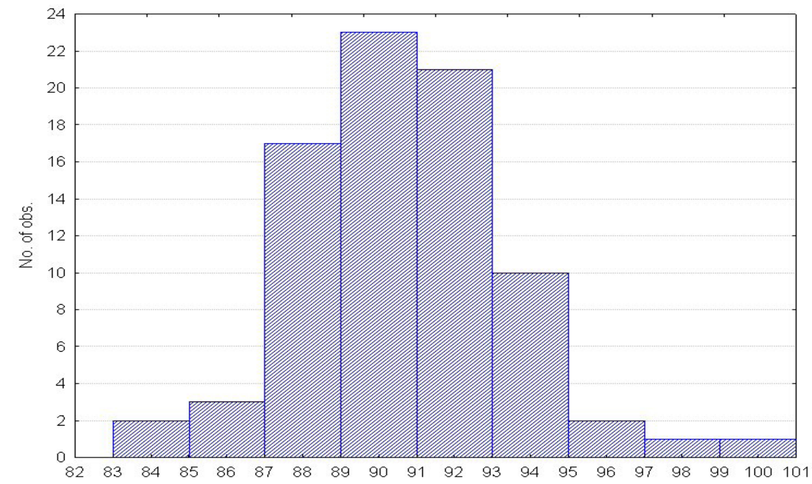


Figure: L'histogramme des fréquences

Représentation graphique

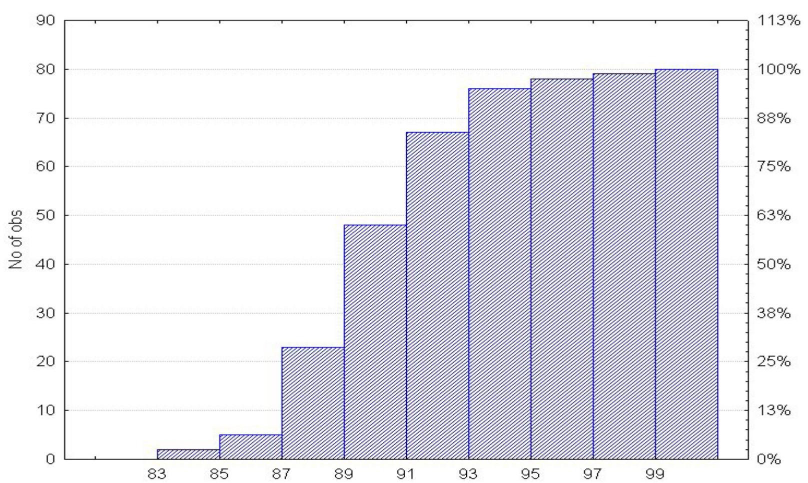


Figure: L'histogramme des fréquences cumulées

Wissem Maazoun MTH2302D: Statistique descriptive 12 / 37

Représentation graphique

Remarque

L'histogramme des fréquences cumulées (ou ogive) permet d'estimer le pourcentage de la population en dessous (ou au dessus) d'une valeur donnée de la variable et inversément.

Exemple

Dans l'exemple portant sur l'indice d'octane, estimer le pourcentage de spécimens de carburant ayant un indice d'octane inférieur à $92,7$.

Représentation graphique

Le diagramme à points

Il s'agit d'une représentation graphique dans laquelle chaque observation est représenté par un point.

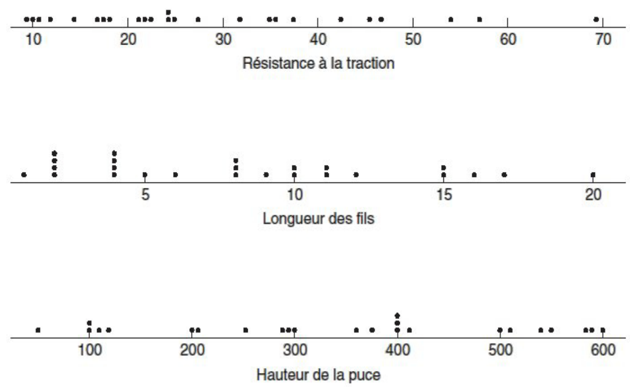
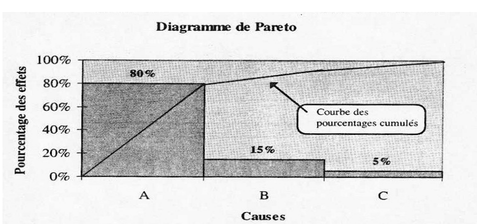


Figure 8.2 Des graphiques de points de la résistance à la traction, de la longueur des fils et de la hauteur de la puce.

Représentation graphique

Le diagramme de Pareto

Le diagramme de Pareto donne, dans l'ordre décroissant, les principales causes à un problème (généralement) liées à la qualité d'une production ou d'un service. Le but du diagramme est de déterminer les quelques causes qui produisent la majorités des effets observés, et facilitant ainsi le choix des mesures correctives à prendre. Il est en effet établi (Pareto) que 20% des causes sont responsables de 80% des effets observés (30% des causes suivantes ne produisent que 15% des effets et le reste des causes, soit 50%, ne produisent que 5% des effets).

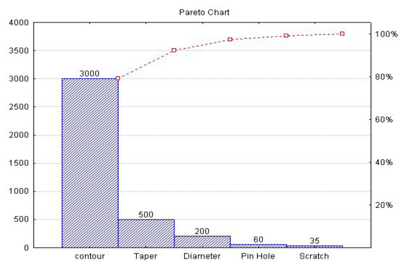


Wissem Maazoun MTH2302D: Statistique descriptive 15 / 37

Représentation graphique

Le diagramme de Pareto (Exemple)

	PROBLEM	LOSS
1	Taper	500.00
2	Diameter	200.00
3	Contour	3000.00
4	Scratch	35.00
5	Pin Hole	60.00



Le diagramme de Pareto

Représentation graphique

Les diagrammes chronologiques

Ce type de diagramme permet de voir, si possible, l'évolution d'une variable en fonction du temps (un titre boursier, la température dans une ville, etc.)

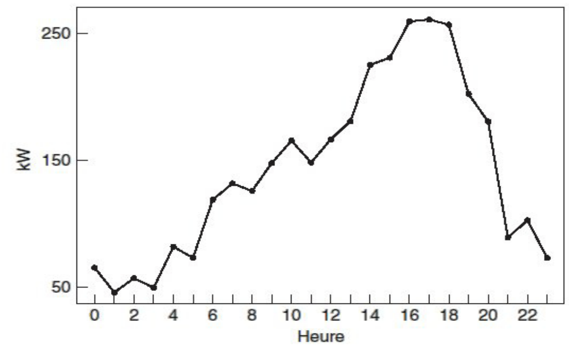


Figure 8.12 La demande horaire d'électricité (en kW) d'un immeuble à bureaux durant la journée de pointe de l'hiver.

Description numérique

Introduction

Il s'agit d'un certain nombre de caractéristiques numériques permettant de décrire des données. On distingue

- les mesures de tendance centrale ;
- les mesures de dispertion (ou variabilité).

Les mesures de tendance centrale

Ce sont des indices qui decrivent le centre d'une distribution de données. On distingue :

- la moyenne ;
- la médiane ;
- le mode.

Wissem Maazoun MTH2302D: Statistique descriptive 18 / 37

La moyenne

La moyenne de l'échantillon est définie par

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

La médiane

La médiane de l'échantillon (notée $\tilde{X}$) est une valeur telle que 50% des observations de l'échantillon lui sont inférieures et 50% lui sont supérieures. Précisément

$$Prop(x_i < \tilde{x}) \leq 50\% \text{ et } Prop(x_i > \tilde{x}) \leq 50\%.$$

Le mode

Le mode de l'échantillon (noté M_0) est la valeur la plus fréquente parmi les données de l'échantillon.

Exemple

On dispose de 6 observations suivantes sur la durée de fonctionnement (en 10^3 heure) d'une marque A d'ampoule électrique :

$$2,6 \quad 3,7 \quad 3,0 \quad 5,5 \quad 2,3 \quad 3,7$$

Déterminer la moyenne, la médiane et le mode de l'échantillon.

Introduction

Les principales mesures de dispersion (ou variabilité) sont :

- l'étendue ;
- l'écart interquartile ;
- la variance ;
- l'écart type ;
- le coefficient de variation.

Soit $x_1, x_2, \ldots, x_n$ un échantillon de n observations d'une population (variable quantitative).

Étendu

L'étendue de l'échantillon (notée R) est définie par

$$R = \max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\} = x_{(n)} - x_{(1)}.$$

Écart interquantile

L'écart interquartile de l'échantillon (noté IQR) est défini par

$$IQR = Q_3 - Q_1 = x_{0,75} - x_{0,25}.$$

Il mesure la dispersion des 50% des données du centre de la distribution.

La variance

Pour exprimer la variance de l'échantillon (notée s^2), on considère d'abord la somme des carrés

$$S_{xx} = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

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

La variance de l'échantillon est définie par

$$s^2 = \frac{S_{xx}}{n-1}.$$

Écart-type

L'écart-type de l'échantillon (noté s) est défini par

$$s = \sqrt{s^2}.$$

Le coefficient de variation

Le coefficient de variation de l'échantillon (noté CV) est défini par

$$CV = \frac{s}{\bar{x}}.$$

Exemple

Calculer la variance, l'écart-type, l'écart interquartile, l'étendue et le coefficient de variation des données de l'exemple précédent.

Wissem Maazoun MTH2302D: Statistique descriptive 24 / 37

Les mesures de dispersion

Quelques utilisations et interprétations de l'écart-type

Soit $x_1, \ldots, x_n$ un échantillon de moyenne $\bar{x}$ et d'écart-type s. Alors

1. Les proportions de données dans les intervalles $\bar{x} \pm s$; $\bar{x} \pm 2s$; $\bar{x} \pm 3s$ doivent être d'environ 68% ; 95% et 99,7% respectivement (en cas de normalité).
2. Pour chaque observation x_i, la valeur standardisée est $z_i = \frac{x_i - \bar{x}}{s}$ (cote Z). Si $z_i < -3$ ou $z_i > 3$, alors x_i constitue une valeur exceptionnelle.

Autres mesures

Les centiles (quantiles ou percentiles)

Soit p un nombre réel entre 0 et 1 ($0 \leq p \leq 1$). On appelle le $100p^{\text{ième}}$ quantile (ou quantile d'ordre p) de l'échantillon $x_1, \ldots, x_n$, le nombre (noté x_p), tel que $100p\%$ des observations sont inférieures à x_p et $100(1-p)\%$ des observations sont supérieures à x_p. Précisément

$$\text{Prop}(x_i < x_p) \leq 100p\% \text{ et } \text{Prop}(x_i > x_p) \leq 100(1-p)\%.$$

Il existe plusieurs méthodes pour déterminer x_p. Parmi celles-ci, l'une des plus utilisées (logiciels statistiques) consiste à

1. Calculer $p \times (n+1) = i + d$, où i est la partie entière de $p \times (n+1)$ et d la partie décimale ($0 \leq p < 1$).
2. Calculer $x_p = x_{(i)} + d \times [x_{(i+1)} - x_{(i)}]$.

Autres mesures

Coefficients d'asymétrie et d'applatissement

Pour un échantillon $x_1, \ldots, x_n$ de taille n d'une variable quantitative,

1. on définit le coefficient d'asymétrie par

$$\hat{\beta}_3 = \frac{n}{(n-1)(n-2)} \times \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}; n > 2.$$

2. on définit le coefficient d'applatissement par

$$\hat{\beta}_4 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \times \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3; \quad n > 3.$$

Autres méthodes d'analyse

Le diagramme de Tukey ("Box-Plot")

Il s'agit d'un graphique constitué d'un rectangle et de deux segments de droite. Les longueurs du rectangle et des segments de droite sont déterminées par les quartiles Q_1, Q_2 et Q_3.

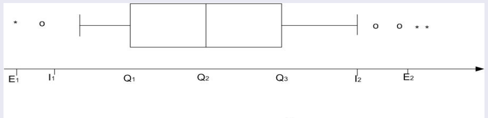


Figure: Schéma d'un digramme de Tukey

I_1 et I_2 sont les limites internes ; E_1 et E_2 sont les limites externes. Ces limites sont définis par

$$I_1 = Q_1 - 1,5IQR; \quad I_2 = Q3 + 1,5IQR;$$
$$E_1 = Q_1 - 3IQR; \quad E_2 = Q3 + 3IQR.$$

Autres méthodes d'analyse

Le diagramme de Tukey ("Box-Plot") (Suite)

Le diagramme permet de détecter la présence de données suspectes et des données aberrantes. Il permet aussi d'évaluer la variabilité et la symétrie des données. En effet

1. la boite donne une idée sur la variabilité des 50% des données du centre ;
2. les segments de droites, en l'absence de données suspectes et aberrantes, donnent une idée sur la variabilité des 25% des données supérieures et des 25% des données inférieures.

Le diagramme permet de comparer plusieurs groupes de données relativement aux caractéristiques mentionnées ci-dessus.

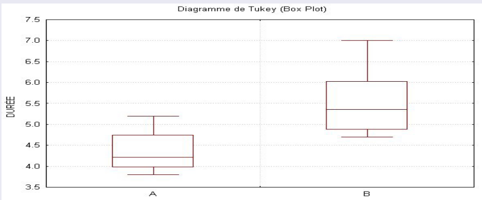
Autres méthodes d'analyse

Exemple

Deux machines (A et B) servent à mettre au point un type de pièce d'équipement. Afin d'étudier le temps nécessaire à la mise au point d'une pièce par chacune des deux machines, on a mesuré les durées requises pour 10 pièces produites par chacune. Les résultats obtenus sont les suivants :

A :	4,9	4,2	4,6	4,1	3,8	5,2	3,9	4,7	4,2	4,0
B :	5,1	4,8	6,0	7,0	5,7	4,9	5,2	5,5	4,7	6,1

Le diagramme de Tukey correspondant est :



Wissem Maazoun MTH2302D: Statistique descriptive 30 / 37

Autres méthodes d'analyse

Le diagramme de dispersion

Le diagramme de dispersion (ou nuage de points) permet de visualiser le type de lien qui existe entre deux variables. Si on étudie la résistance (Y) d'un béton en fonction du temps (durée) de séchage (X), l'expérience consiste à obtenir des données de la forme (x, y) de façon expérimentale, où x désigne une durée de séchage fixée et y la résistance observée. Lorsqu'une telle expérience est répétée un certain nombre de fois, on obtient des observations de la forme $(x_i, y_i), i = 1, \ldots, n$. La reperésentation des n points $(x_1, y_1)(x_2, y_2), \ldots (x_n, y_n)$ dans le plan constitue un diagramme de dispersion (ou un nuage de points).
Un tel diagramme peut indiquer un tendance linéaire, exponentielle ou autre, ou bien simplement aucune tendance.

Autres méthodes d'analyse

Exemple

Au cours d'une expérience visant à évaluer la performance d'un modèle de véhicule automobile, les données suivantes, portant sur le nombre de litres de carburant (x) et la distance (en km) parcourue (y), ont été recueillies.

Nombre de litres (x)	Distance (y)
34,3	450,2
29,2	410,5
24,5	354,1
33,8	472,5
26,2	365,8

Autres méthodes d'analyse

Exemple

Le diagramme de dispersion pour ces données est de la forme suivante :

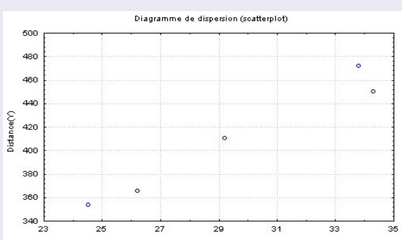


Figure: Graphique du diagramme de dispersion.

Le diagramme montre qu'il existe une relation entre le nombre de litres de carburant et la distance parcourue. Nous verrons au chapitre 13 comment déterminer une telle relation.

Le coefficient de corrélation

Il s'agit d'un indice qui permet de mesurer le degré d'association linéaire entre deux variables. Le nuage de points donne un aperçu du lien, tandis que le coefficient de corrélation permet de le quantifier.
Pour n observations (sur deux variables) de la forme $(x_i, y_i), i = 1, \ldots, n$, on définit le coefficient de corrélation par

$$r = \frac{S_{xy}}{(S_{xx} \times S_{yy})^{\frac{1}{2}}}$$

$$\text{avec} \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ et } S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Le coefficient de corrélation (suite)

C'est un nombre sans unité et on a toujours : $-1 \leq r \leq 1$.
Plusieurs possibilités :

1. si $r = -1$ ou $r = 1$. Les points du diagramme de dispersion sont tous sur la droite. On dit alors qu'on a une corrélation parfaite ;
2. si $r = 0$, on a une abscence de corrélation (mais cela ne veut pas dire qu'il n'y a pas de lien entre les deux variables). Sur le diagramme de dispersion, les points sont dispersés au hasard ;
3. dans les autres cas, la corrélation peut être forte, moyenne ou faible. De plus, si $r > 0$, on a une corrélation positive ; i.e. les deux variables varient dans le même sens. Et si $r < 0$, on a une corrélation négative et les deux variables varient en sens contraire.

Les données groupées

Il arrive parfois qu'un échantillon $x_1, \ldots, x_n$ soit en fait constitué de p valeurs distinctes $x_1, \ldots, x_p$, où chaque valeur x_j est répétée n_j fois, pour $j = 1, \ldots, p$. Des données de cette nature sont souvent présentées dans un tableau de la forme suivante

valeurs (x_j)	x_1	x_2	$\ldots$	x_p
effectif (n_j)	n_1	n_2	$\ldots$	n_p

On a alors $\sum_{j=1}^{p} n_j = n$; $\quad \sum_{i=1}^{n} x_i = \sum_{j=1}^{p} n_j x_j$;

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{j=1}^{p} n_j x_j}{\sum_{j=1}^{n} n_j}$$

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

Wissem Maazoun MTH2302D: Statistique descriptive 36 / 37

Exemple

Soit les 50 observations suivantes

(x_j)	60	61	62	63	64	65	66
(n_j)	2	8	15	14	6	4	1

Déterminer $\bar{x}$, s^2 et $\tilde{x}$.