8. Statistique descriptive

MTH2302D

S. Le Digabel, École Polytechnique de Montréal

A2017

(v1)

Plan

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 une interprétation des données afin de pouvoir prendre des décisions.
- La statistique utilise
  - des notions de probabilités,
  - des notions de mathématiques.

Introduction (suite)

Définition

La statistique descriptive est 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.

Terminologie

- L'*univers* est l'ensemble des objets sur lesquels porte l'étude statistique.
- Une *variable* est une caractéristique selon laquelle l'univers est étudié.
- La *population* est l'ensemble de toutes les *mesures* ou *observations* de la variable dans l'univers considéré.
- Une *unité expérimentale* est un objet de l'univers, sur lequel la variable est mesurée.
- Un *échantillon* est un sous-ensemble
  - de l'univers : s'il est composé d'unités exérimentales,
  - de la population : s'il est composé de mesures de la variable.

Terminologie (suite)

- Un *paramètre* est une mesure caractérisant la variable **dans la population**.

  Par exemple : la moyenne de la population.

  En général, la vraie valeur d'un paramètre est inconnue.

- Une *statistique* est une mesure caractérisant la variable **dans un échantillon** de la population.

  Par exemple : la moyenne échantillonnale.

  Une statistique peut être calculée.

Exemple 1

On a mesuré l'indice d'octane de 80 spécimens de carburant et obtenu les résultats du tableau suivant :

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

Exemple 2

On a tiré 25 circuits électroniques de la production d'une usine et on a mesuré la longueur et la résistance à la traction des fils d'interconnexion de chaque circuit.

No. de l'observation	Résistance à la traction (y)	Longueurs des fils (x)
1	9.95	2
2	24.45	8
3	31.75	11
4	35.00	10
5	25.02	8
6	16.86	4
7	14.38	2
8	9.60	2
...	...	...

Utilité des descriptions graphiques

- Présenter les données de façon à en avoir une vue d'ensemble.
- Utile pour interpréter les données et observer facilement :
  - tendance centrale,
  - étalement,
  - comparaison,
  - valeurs suspectes ou aberrantes,
  - ...

Distribution de fréquences

- L'ensemble des valeurs mesurées de la variable est subdivisé en sous-intervalles (*classes*). Si on a n données, environ $\sqrt{n}$ classes est un bon choix.
- On construit un tableau de la forme :

Classe	Fréquence	Fréquence cumulative	Pourcentage	Pourcentage cumulatif
$a \leq x \leq b$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\vdots$				

Exemple 3

On a mesuré l'indice d'octane de 80 spécimens de carburant et obtenu les résultats du tableau présenté plus haut. Le tableau de fréquences obtenu avec `Statistica` est :

Category	Frequency table: Octane1 (octane.sta) K-S d=,08357, p> .20; Lilliefors p<,20					
	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
80,00000<x<=85,00000	2	2	2,50000	2,5000	2,50000	2,5000
85,00000<x<=90,00000	31	33	38,75000	41,2500	38,75000	41,2500
90,00000<x<=95,00000	43	76	53,75000	95,0000	53,75000	95,0000
95,00000<x<=100,0000	3	79	3,75000	98,7500	3,75000	98,7500
100,0000<x<=105,0000	1	80	1,25000	100,0000	1,25000	100,0000
Missing	0	80	0,00000		0,00000	100,0000

Histogramme

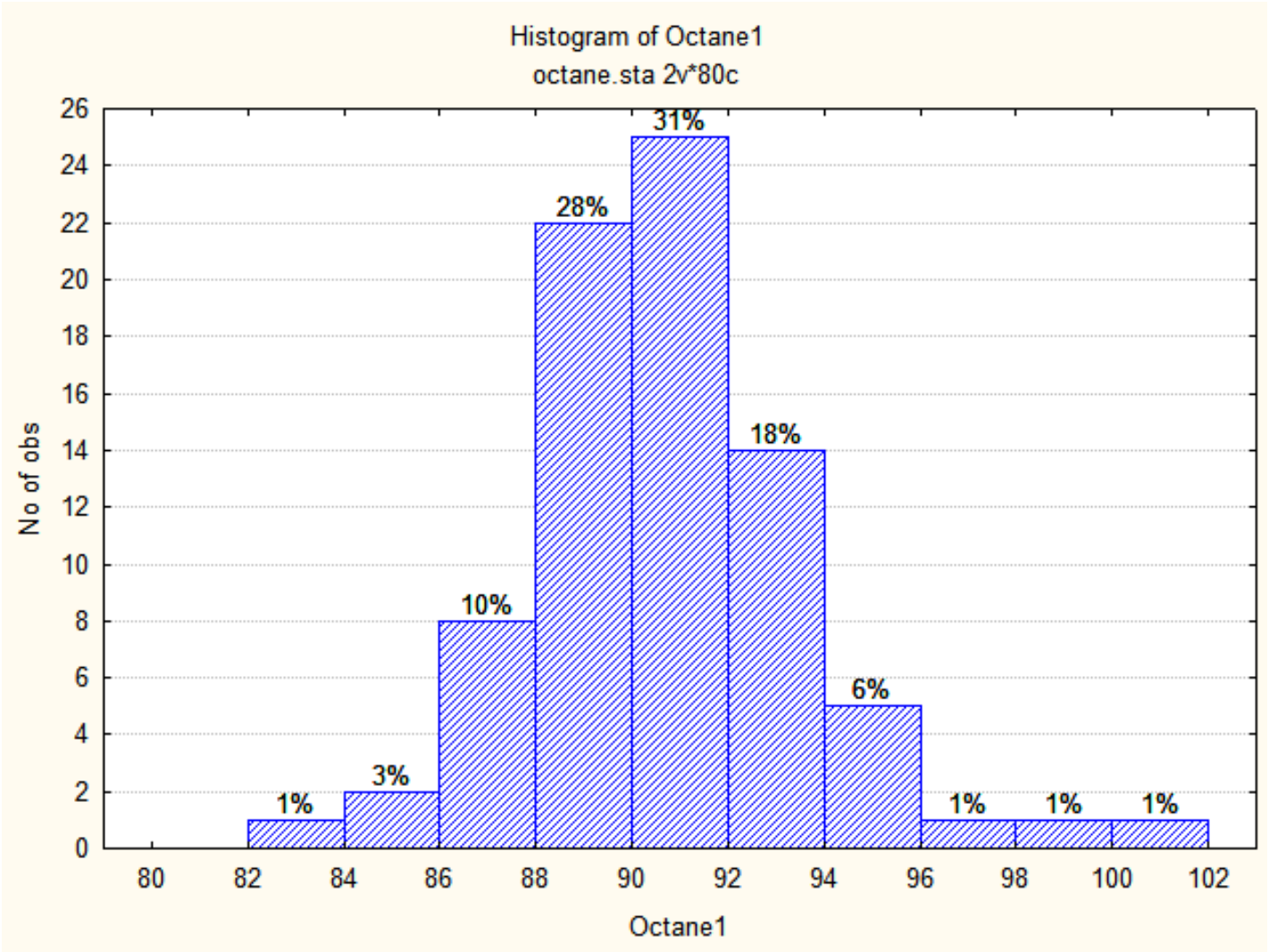
- L'ensemble des valeurs observées est subdivisé en intervalles (*classes*). Si on a n données, environ $\sqrt{n}$ sous-intervalles est un bon choix. Les intervalles ne sont pas nécessairement égaux.
- Sur chaque intervalle on construit un rectangle dont l'aire est proportionnelle à la fréquence relative de la classe.

  Si les intervalle sont égaux alors la hauteur du rectangle est la fréquence de la classe correspondante.
- On peut avoir un histogramme
  - Des fréquences.
  - Des fréquences cumulées.

Exemple 4

On a mesuré l'indice d'octane de 80 spécimens de carburant et obtenu les résultats du tableau présenté plus haut. Les histogrammes des fréquences et des fréquences cumulées obtenus avec `Statistica` sont aux pages suivantes.

Exemple 5


Histogram of Octane1
octane.sta 2v*80c
No of obs
Octane1
0
2
4
6
8
10
12
14
16
18
20
22
24
26
80
82
84
86
88
90
92
94
96
98
100
102
1%
3%
10%
28%
31%
18%
6%
1%
1%
1%

Exemple 6

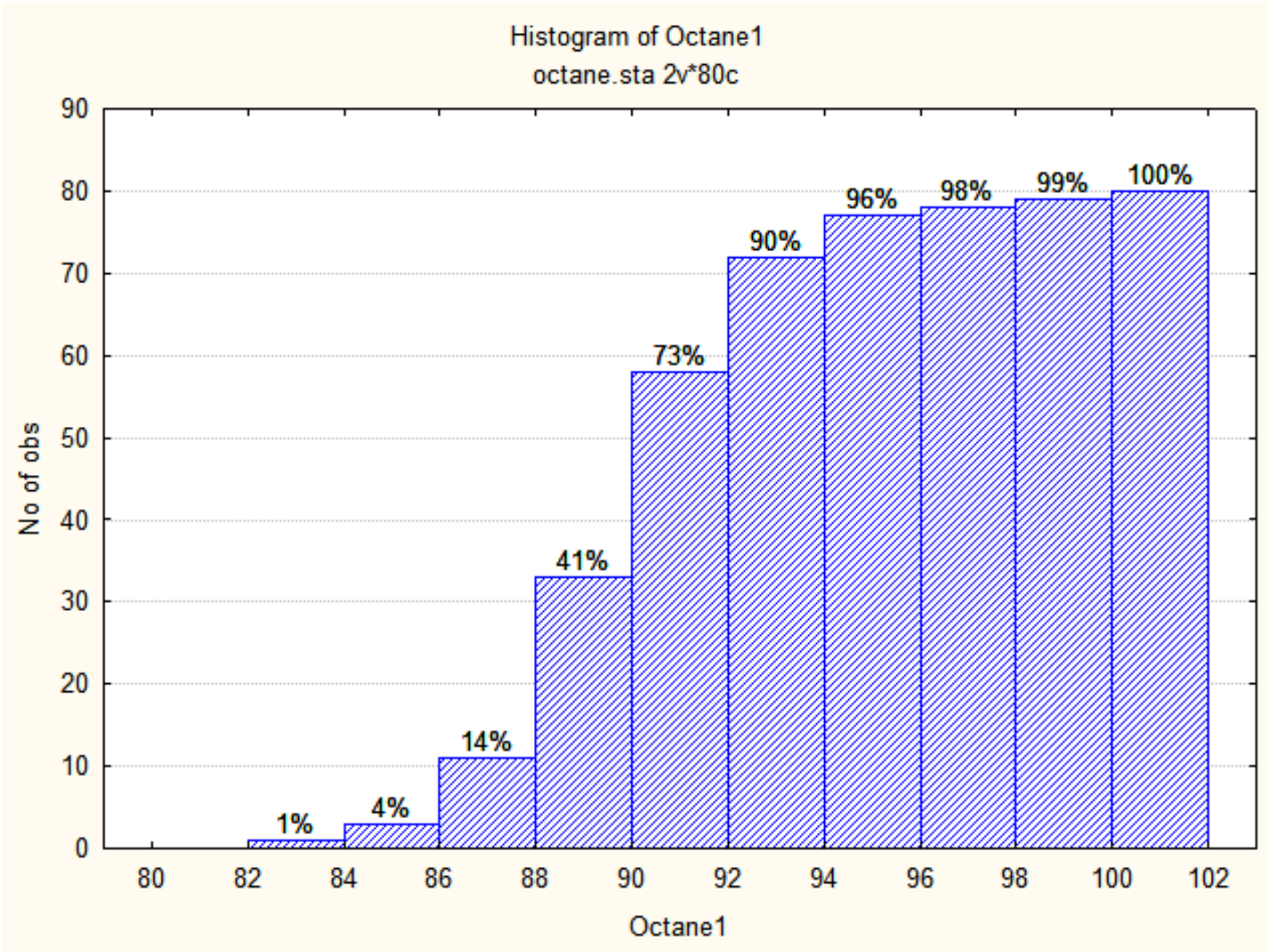

Histogram of Octane1
octane.sta 2v*80c
No of obs
Octane1
0
10
20
30
40
50
60
70
80
90
80
82
84
86
88
90
92
94
96
98
100
102
1%
4%
14%
41%
73%
90%
96%
98%
99%
100%

Diagramme tige-feuille

- Chaque valeur observée est divisée en deux parties : les premiers chiffres (tige) et les chiffres restants (feuille).
- On arrange les données dans un tableau où chaque ligne commence par une tige, suivie des feuilles correspondant à cette tige, dans l'ordre croissant.
- Avantage : les donnés individuelles sont toujours visibles.

Exemples

Exemple 7

les données

10.2 11.5 11.9 13.1 10.2 12.4 12.6 11.6 10.7 13.2

donnent

tige	feuilles	effectifs
10	2.2.7	3
11	5.6.9	3
12	4.6	2
13	1.2	2

Exemple 8

On a mesuré l'indice d'octane de 80 spécimens de carburant et obtenu les résultats du tableau présenté plus haut. Le diagramme tige-feuille obtenu avec `Statistica` est à la page suivante.

	Stem and Leaf Plot: Octane1 (c Octane1 one leaf=1 case	
stem°leaf (leaf unit=1,000000, e.g., 6°5 = 6,500000)	Class n	Percentiles
80°	0	
81°	0	
81°	0	
82°	0	
82°	0	
83° 4	1	
83°	0	
84° 3	1	
84°	0	
85° 3	1	
85°	0	
86°	0	
86° 77	2	
87° 4	1	
87° 56789	5	
88° 23334	5	
88° 556679	6	25%
89° 0233	4	
89° 678899	6	
90° 01133444	8	median
90° 56789	5	
91° 00011122	8	
91° 56688	5	
92° 2223	4	75%
92° 6777	4	
93° 02334	5	
93° 7	1	
94° 224	3	
94° 7	1	
95°	0	
95° 6	1	
96° 1	1	
96°	0	
97°	0	
97°	0	
98°	0	
98° 8	1	
99°	0	
99°	0	
100° 3	1	
100°	0	
min = 83,40000 max = 100,3000 Total N:	80	

Diagramme à points, nuage de points

- Diagramme à points : **une seule variable**
  - Chaque observation est représentée par un point au-dessus de la valeur correspondante sur l'axe horizontal.
  - S'il y a plus d'une observation pour une valeur donnée, on superpose les points.
- Nuage de points : **deux variables**
  - Les deux axes correspondent aux valeurs des deux variables.
  - Chaque couple d'observations est représenté par un point dans le plan.
  - S'il y a plus d'une donnée avec les même valeurs, le point correspondant peut être représenté par un autre symbole.
  - Permet de visualiser les relations possibles entre les variables.
- Nuage de points : trois variables. Ici, chaque point dans l'espace correspond à un triplet d'observations.

Exemple 9

On a tiré 25 circuits électroniques de la production d'une usine et on a mesuré la longueur et la résistance à la traction des fils d'interconnexion de chaque circuit. Le nuage de points pour ces deux variables, obtenu avec `Statistica` est :

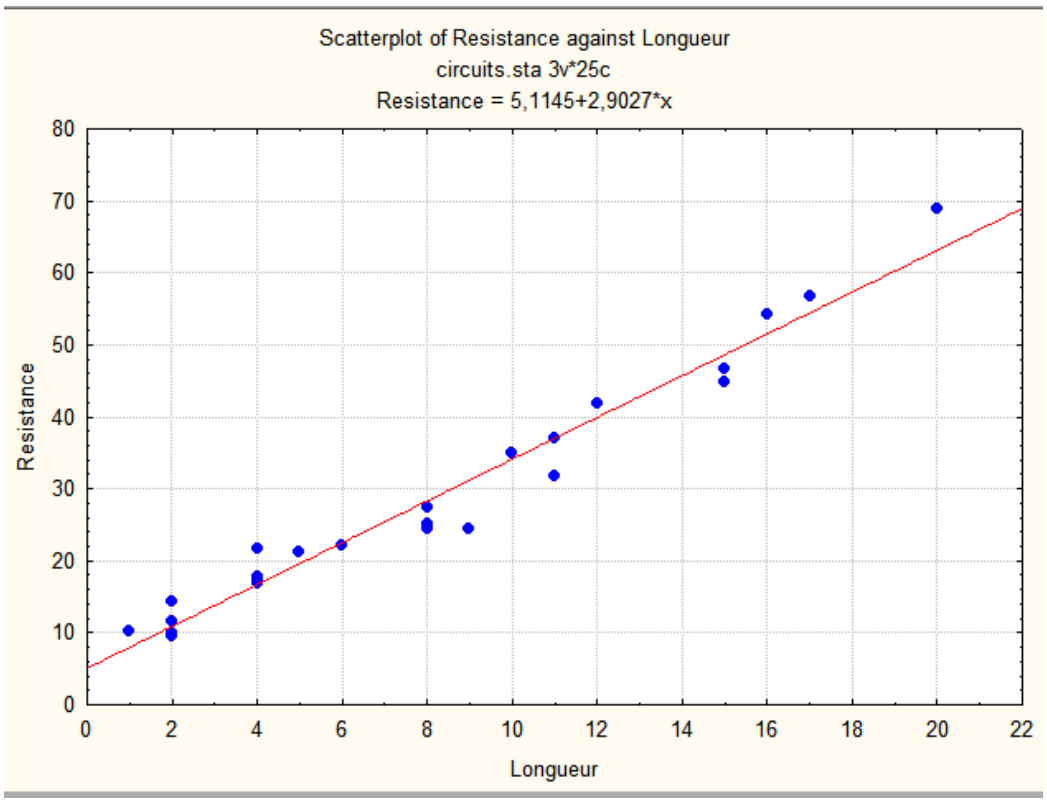


Diagramme en boîte

Définition

Un *quartile* est l'une des trois valeurs, dénotées Q_1, Q_2 et Q_3, qui divisent les données en quatre parties égales, de sorte que chaque partie contienne le quart des données.

Diagramme en boîte, ou *boîte à moustaches*, ou *boîte de Tukey*, ou *box & whisker plot*, ou *box plot* :

- Une boîte est dessinée, centrée sur le deuxième quartile avec deux côtés alignés avec les premier et troisième quartiles.
- Un segment est dessiné de chaque côté de la boîte, l'un jusqu'à la valeur minimum des données, l'autre jusqu'à la valeur maximum.
- Utile pour comparer deux échantillons.

Exemple 10

On a mesuré l'indice d'octane de **deux** échantillons de 80 spécimens de carburant. Le diagramme en boîte obtenu avec `Statistica` est :

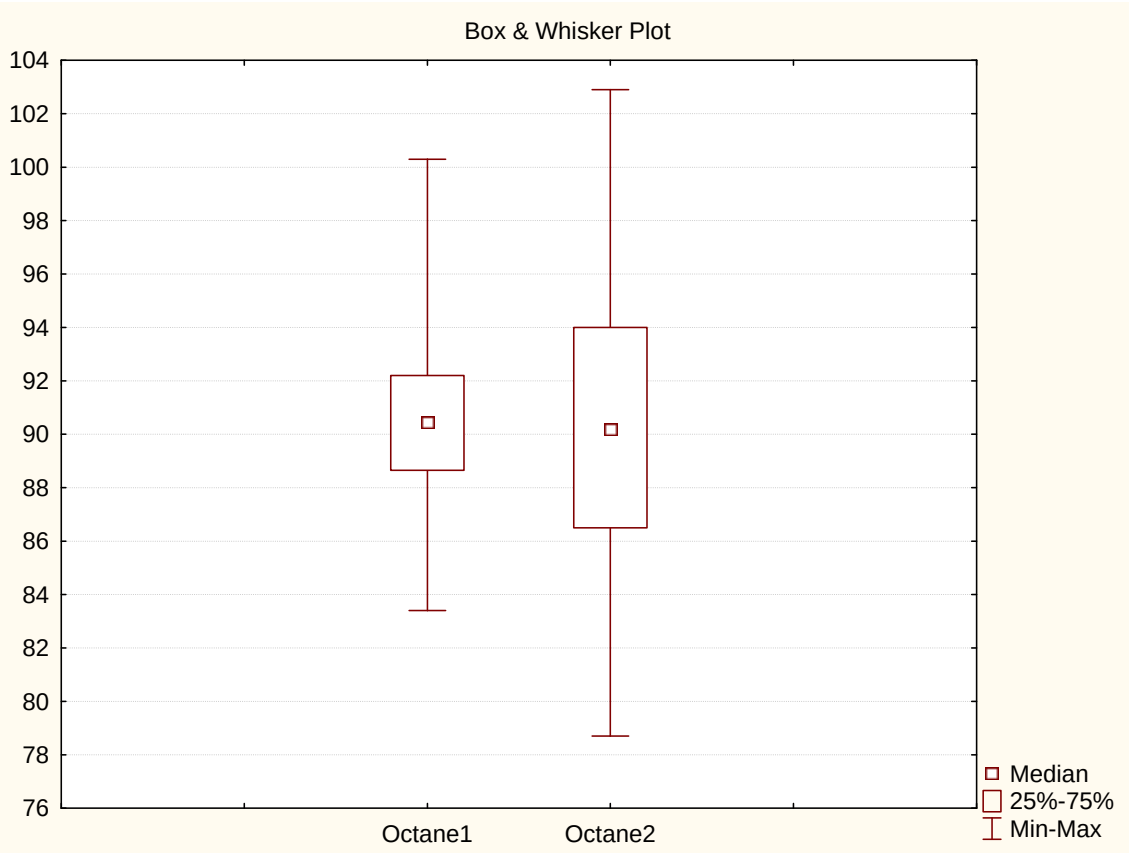


Diagramme de Pareto

- En abscisse : les catégories (possiblement non numériques), en ordre décroissant d'effectifs.
- En ordonnée : la fréquence (effectif) de la catégorie.
- Pour chaque catégorie, on trace un rectangle dont la hauteur est l'effectif de la catégorie.
- On relie les valeurs des effectifs cumulés pour obtenir un graphe linéaire par morceaux.
- Permet de représenter des catégories non numériques.
- Permet de visualiser rapidement les catégories les plus fréquentes.

Exemple 11

On distingue les défauts suivants pour une pièce métallique faisant partie d'une portière d'automobile :

Défaut	Effectif
Forme déficiente	30
Bosses, creux, rainures	4
Absence de lubrifiant	5
Mauvais détourage	21
Mauvais ordre	6
ébarbage non effectué	5
Fentes ou trous manquants	6
Autre défaut	4

Exemple 11 (suite)

Le diagramme de Pareto obtenu avec Statistica est :

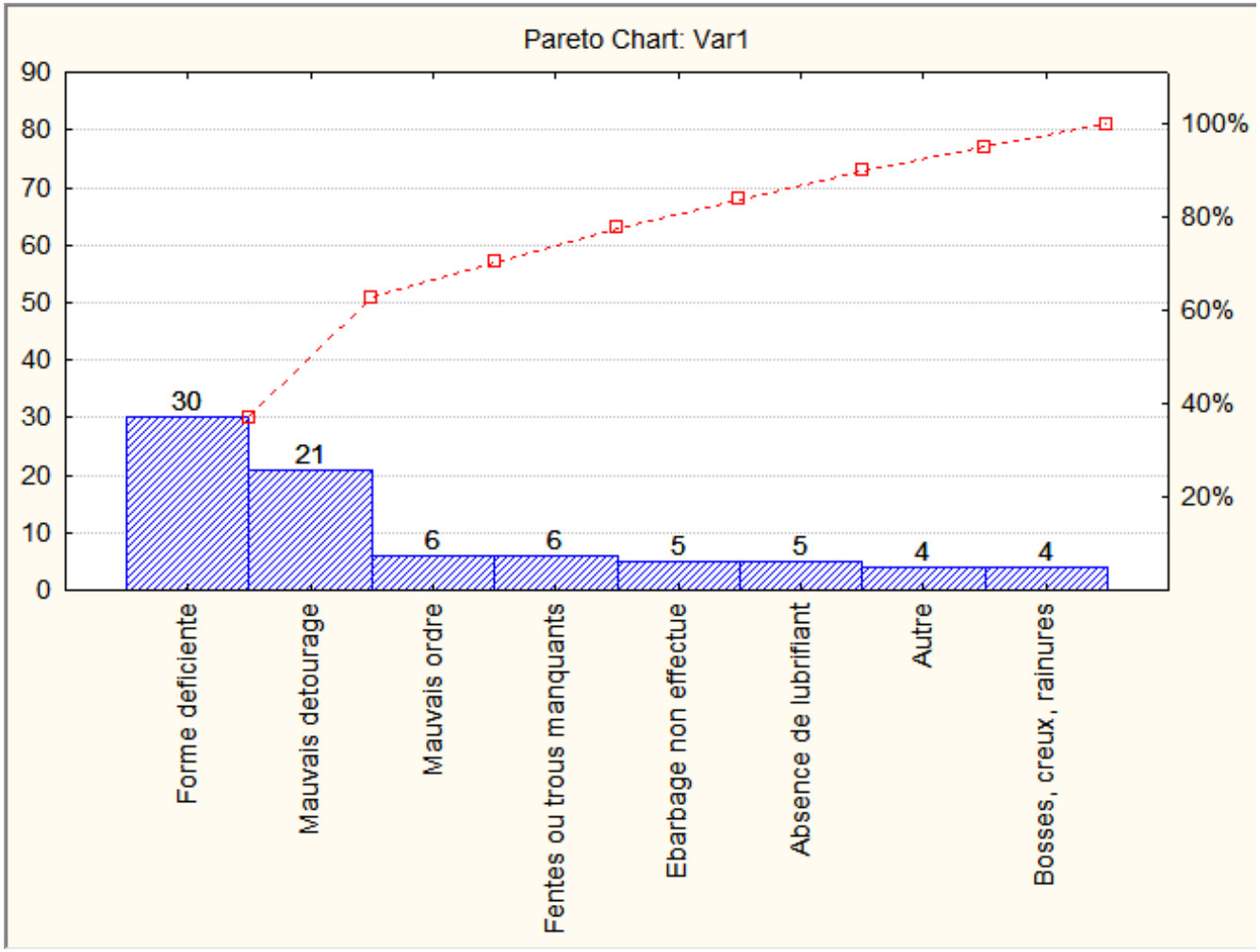
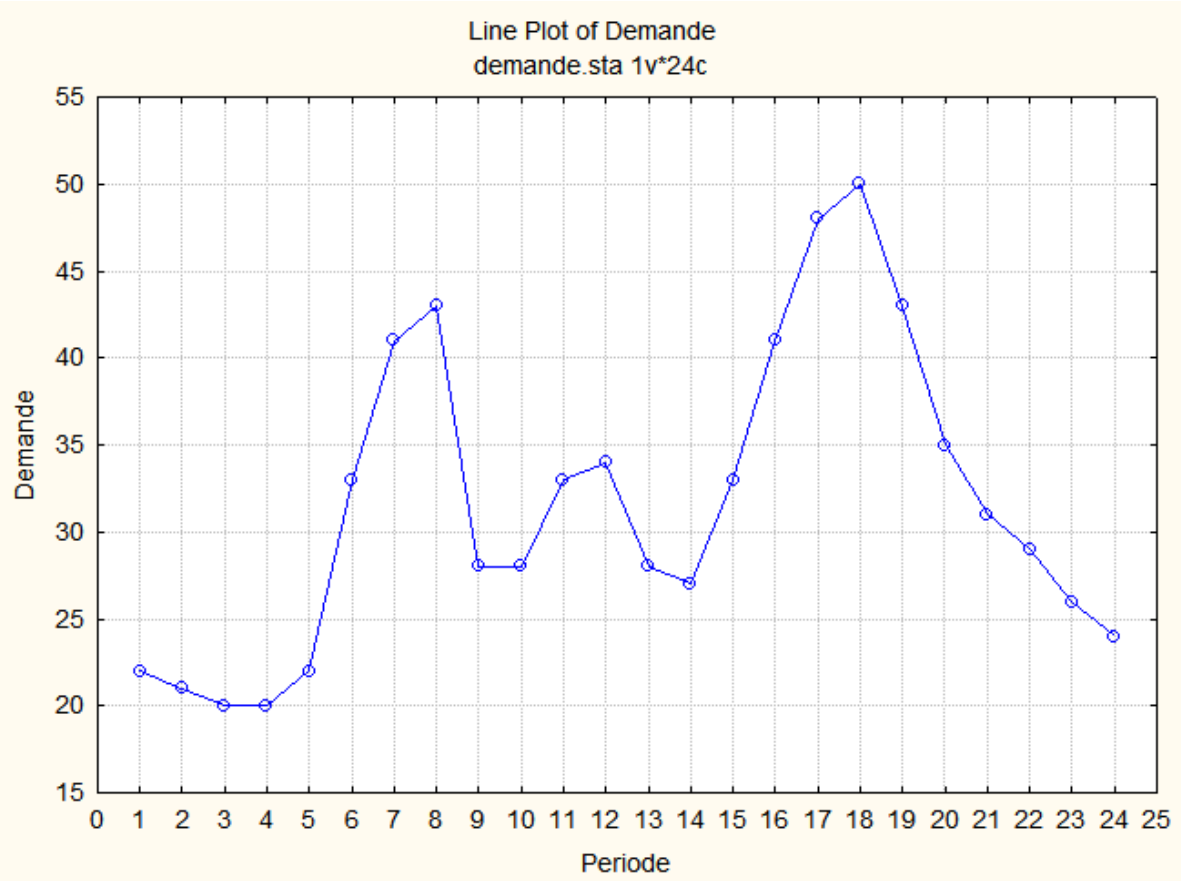


Diagramme chronologique

- Courbe linéaire par morceaux.
- En abscisse : le temps.
- En ordonnée : la valeur des observations pour chaque période de temps.
- Utile pour observer l'évolution de la variable dans le temps.

Exemple 12

La demande en électricité d'une région a été mesurée à chaque heure sur une période de 24 heures. Le diagramme chronologique de la demande, obtenu avec `Statistica` est :



Équivalences de terminologie

		Statistica
distribution de fréquences		frequency table
histogramme	diagramme en barres	histogram
diagramme tige-feuille	histogramme de Tukey	stem and leaf plot
diagramme à points		
nuage de points	diagramme de dispersion	scatterplot
diagramme en boîte	diagramme de Tukey	box and whisker plot
diagramme de Pareto	$\simeq$ polygone d'effectifs	Pareto chart
diagramme chronologique		$\simeq$ line plot

Trois types de mesures numériques

- Mesures de tendance centrale : moyenne, médiane, mode.
- Mesures de dispersion (étalement) : étendue, écart interquartile, variance, écart-type, coefficient de variation, centiles.
- Mesure d'association : coefficient de corrélation.

Tendance centrale : moyenne, médiane, mode

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

La *moyenne* de l'échantillon, ou *moyenne échantillonnale* est

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

La moyenne n'est pas nécessairement égale à la valeur d'une des données.

Tendance centrale : moyenne, médiane, mode (suite)

La *médiane* de l'échantillon, dénotée $\tilde{x}$, est une valeur telle que 50% des observations lui sont supérieures et 50% lui sont inférieures.

Si $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ sont les données en ordre croissant alors

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ est impair} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{si } n \text{ est pair.} \end{cases}$$

Si n est impair alors la médiane est égale à l'une des données. Si n est pair, elle n'est pas forcément égale à l'une des données.

Tendance centrale : moyenne, médiane, mode (suite)

- Le *mode* de l'échantillon est la valeur la plus fréquente des données.

  Un échantillon peut avoir plusieurs modes.

  Le mode est nécessairement égal à l'une des données.

- On peut aussi définir le mode comme le point milieu de la classe ayant le plus grand effectif.

Dispersion : étendue, écart interquartile

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

- L'*étendue* de l'échantillon est

$$R = \max\{x_1, x_2, \ldots, x_n\} - \min\{x_1, x_2, \ldots, x_n\}.$$

- L'*écart interquartile* est

$$\text{IQR} = Q_3 - Q_1$$

où Q_1 et Q_3 sont les premier et troisième quartiles.

Dispersion : étendue, écart interquartile (suite)

Méthode pour le calcul des quartiles

1. Utiliser la médiane pour diviser les données en deux parties égales. Ne pas inclure la médiane dans les deux sous-ensembles obtenus.

   Poser : $Q_2 =$ médiane de l'échantillon.

2. Poser
   $Q_1 =$ médiane du sous-ensemble des valeurs inférieures à Q_2.
   $Q_3 =$ médiane du sous-ensemble des valeurs supérieures à Q_2.

Dispersion : variance, écart-type, coeff. de variation

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

La *variance* de l'échantillon, dénotée s^2, est définie par

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

avec

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

Dispersion : variance, écart-type, coeff. de variation (suite)

- L'*écart-type* de l'échantillon est $s = \sqrt{s^2}$.
- Le *coefficient de variation* de l'échantillon mesure la dispersion relative des données autour de la moyenne :

$$\text{CV} = s/\overline{x} \ .$$

Moments (centrés) :

- $\hat{\mu}_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^k$ ou $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^k$.
- $\hat{\beta}_1 = \hat{\mu}_3 / s^3$ (asymétrie).
- $\hat{\beta}_2 = \hat{\mu}_4 / s^4 \quad (-3)$ (aplatissement).

Exemple 13

Avec les données :

115	2456	534	3915	1046	1916	1117	1303	865	340
575	3563	4413	500	2096	149	1511	2244	695	1021

- Donner le tableau de fréquences avec cinq classes de largeur 1000.
- Calculer R, $\overline{x}$, $\tilde{x}$, s^2, les quartiles, IQR, le mode, et CV.

Dispersion : centiles

Soit $0 < p < 1$. Le $(100 \cdot p)^e$ *centile* de l'échantillon est un nombre x_p tel que :

- $100p$ % des données sont inférieures à x_p.
- $100(1-p)$ % des données sont supérieures à x_p.

Association : coefficient de corrélation

Soit n observations de deux variables quantitatives (x_i, y_i) avec $i = 1, 2, \ldots, n$. Le *coefficient de corrélation* de x et y est

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

où

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

On peut montrer que $-1 \leq r \leq 1$.

Interprétation du coefficient de corrélation

- Si $|r| = 1$ alors il y a corrélation parfaite entre les x_i et les y_i. Les points du diagramme de dispersion sont tous sur une même droite.
- Si $r = 0$ alors il n'y a pas de corrélation entre les x_i et les y_i. Les points du diagramme de dispersion sont distribués "au hasard" dans le plan.
- Si $-1 < r < 1$ alors il y corrélation forte, moyenne ou faible entre les x_i et les y_i.

  La tendance des points du diagramme de dispersion à former une droite dépend de r.

  Si $r > 0$ alors les variables x et y varient dans le même sens (corrélation positive).

  Si $r < 0$ alors les variables x et y varient en sens opposé (corrélation négative).

Exemple 14

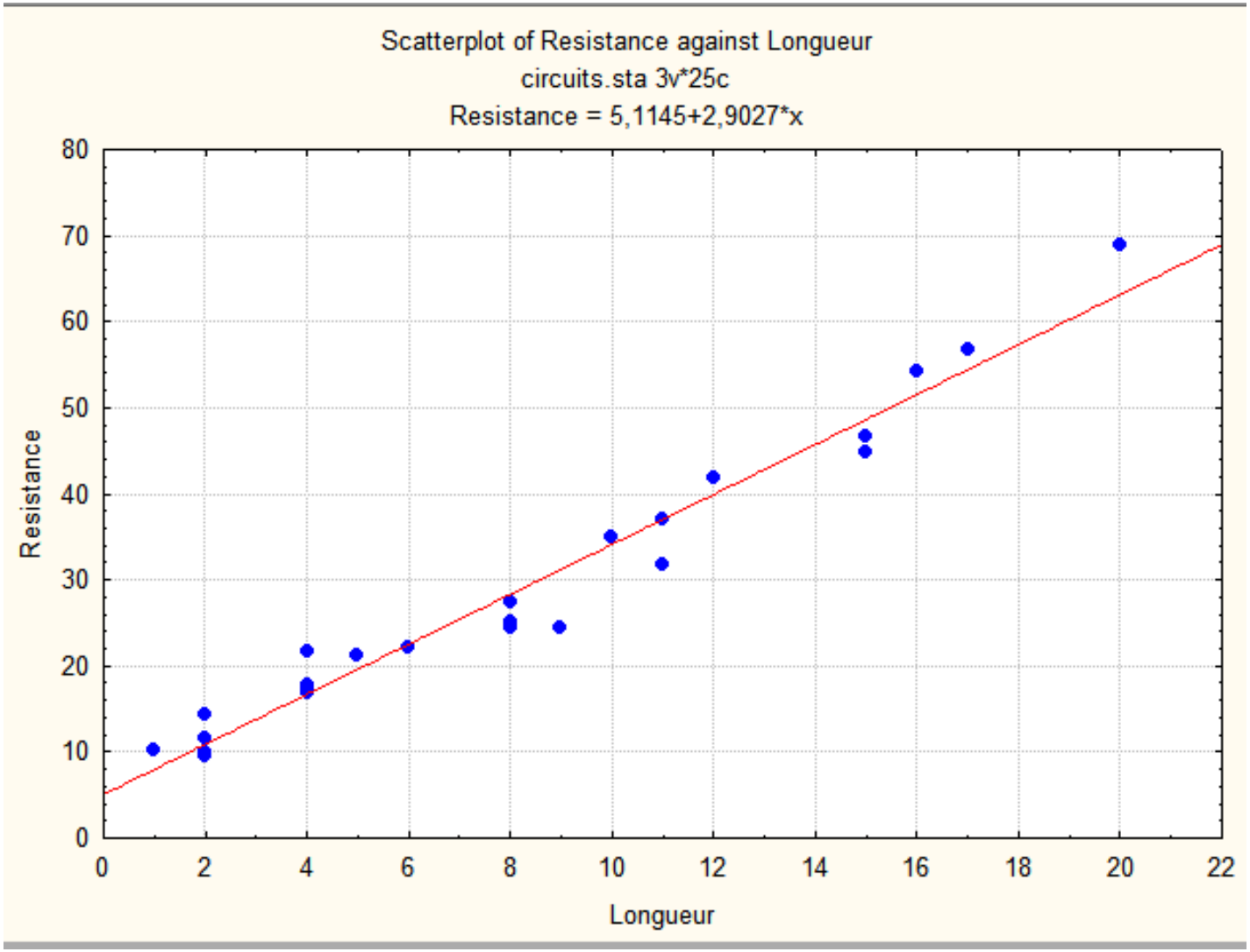
On a tiré 25 circuits électroniques de la production d'une usine et on a mesuré la longueur (x) et la résistance à la traction (y) des fils d'interconnexion de chaque circuit.

Le coefficient de corrélation calculé par `Statistica` est 0.9818.

Il y a donc une forte corrélation positive entre la longueur et la résistance.

Ceci est illustré par le fait que le nuage de points correspondant aux variables x et y est très proche d'une droite.

Exemple 14 (suite)



Données groupées

Supposons qu'un échantillon $x_1, x_2, \ldots, x_n$ est constitué de p valeurs distinctes $x_{(1)}, x_{(2)}, \ldots, x_{(p)}$ $(p < n)$, où chaque valeur $x_{(j)}$ est répétée n_j fois.

On peut présenter cet échantillon dans un tableau

Valeur $(x_{(j)})$	$x_{(1)}$	$x_{(2)}$	$\cdots$	$x_{(p)}$
Effectif (n_j)	n_1	n_2	$\cdots$	n_p

Les formules pour la moyenne, variance, etc. peuvent être réécrites en fonction des $x_{(j)}$ et des n_j.

Si au lieu de valeurs individuelles on a des classes sous forme d'intervalles (tableau de fréquences) alors on utilise les points milieu des intervalles pour approximer les $x_{(j)}$ dans le calcul de $\overline{x}$, s^2, etc.