Polytech Nancy **2020 / 2021**
3e année

Statistiques

Introduction

Chapitre 1 – Statistique descriptive

Marie-Christine SUHNER

marie-christine.suhner@univ-lorraine.fr

1.1 Introduction

La statistique

La statistique regroupe l'ensemble des méthodes scientifiques à partir desquelles on recueille, organise, présente et analyse des données dans l'objectif d'en **tirer des conclusions et de prendre des décisions** en situation concrète soumise aux aléas de l'incertain

Contexte d'incertitude

- Echantillonnage
- Absence de contrôle parfait sur les processus
- Erreurs de mesure
- Variations dans les matières premières…

Branches de la statistique

- Collecte des données
- **Description des données collectées**

  Statistique descriptive : présentation, représentation graphique, calcul de caractéristiques numériques
- **Interprétation des résultats sur une population globale**

  Modélisation probabiliste : ajustement d'une loi de probabilité théorique sur les données

  Inférence statistique : raisonnement inductif, passage du particulier au général
- Construction de modèles (régression et prédiction, plans d'expérience et analyse de variance)
- *Analyse de données (analyses factorielles et classification)*

1.1 Principales étapes d'une étude statistique

❖ Identification du problème

Définition des objectifs et des variables du problème

- Tirer des conclusions d'une portée générale à partir d'informations incomplètes et particulières
- Aider à prendre des décisions judicieuses dans un contexte d'incertitude

Besoins en informations

Hypothèses

❖ Détermination du cadre de l'étude

Choix de la méthode d'analyse

Choix de la méthode de collecte des données

Choix des instruments de mesure

❖ Recueil des données

Elaboration du plan d'échantillonnage

Collecte des données

❖ Traitement des données

Codification

Saisie et vérification

Imputation (traitement des données manquantes, invalides ou incomplètes)

Contrôle de la qualité des données

Mise en œuvre de la méthode choisie d'analyse des données

Production des résultats (informations)

❖ Interprétation des informations et conclusions

Synthèse des informations

Analyse des résultats sur l'échantillon et interprétation pour la population

Justification des conclusions et recommandations

1.1 Définitions

❖ Caractère (ou variable)

Caractéristique faisant l'objet de l'étude

Nature du caractère :

caractère qualitatif — caractère quantitatif

❖ Individu (ou unité statistique)

Entité de base appartenant à la population ou à l'échantillon sur laquelle on mesure ou on observe le caractère

❖ Population

Ensemble des individus sur lesquels on pourrait faire la mesure ou l'observation

Taille de la population : N

❖ Echantillon

Partie de la population sur laquelle sont effectuées les mesures ou les observations

Taille de l'échantillon : n

échantillon exhaustif (sans remise)

échantillon non exhaustif (avec remise)

❖ Remarques

- Une population finie sur laquelle on effectue un échantillonnage non exhaustif peut être considérée comme infinie
- Un échantillonnage exhaustif réalisé sur une population très grande (n très petit devant N) est considéré comme non exhaustif

1.2 Statistique descriptive à une dimension

- *Traitement d'une seule variable :*
  répartition des mesures ou observations, synthèse et analyse
- *Données d'observations, expérimentales, séries statistiques ou résultats d'un sondage*

❖ Mise en ordre des données

– Rangement des données

Tri par ordre croissant ou décroissant

– Fréquences absolues (ou effectifs)

fréquences absolues pour chaque valeur, classe ou modalité

n_i

– Fréquences relatives

fréquences relatives pour chaque valeur, classe ou modalité

$$f_i = \frac{n_i}{\sum_i n_i}$$

– Caractère qualitatif

Etat d'une pièce x_i	Nombre de pièces n_i
Bon état	27
Réparable	15
Irréparable	2

– Caractère quantitatif

- discontinu

Nombre d'erreurs d'assemblage x_i	Nombre d'appareils n_i	Fréquence cumulée
0	101	0,27
1	140	0,64
2	92	0,89
3	42	1,00

1.2 Statistique descriptive à une dimension

❖ Mise en ordre des données (suite)

– Caractère quantitatif

- continu

Durée de vie x_i	Nombre d'appareils n_i
Moins de 2000	2
$2000 \le X < 2100$	12
$2100 \le X < 2200$	24
$2200 \le X < 2300$	9
2300 et plus	3

– nombre de classes : q (de préférence entre 5 et 20)

Pour $\sum_i n_i$ grand $\quad q \cong \sqrt{\sum_i n_i}$

Formule de Sturges $\quad q \cong 1 + 3{,}322 \log \sum_i n_i$

– souvent, même amplitude pour chaque classe

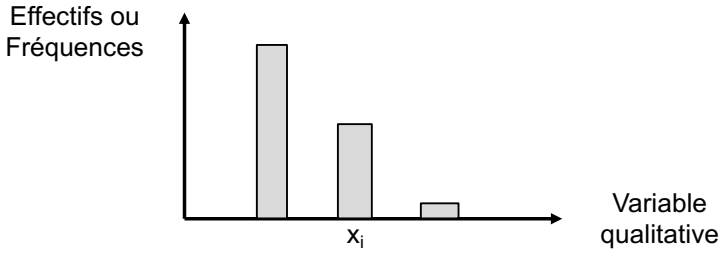
E / q

avec l'étendue $E = x_{Max} - x_{min}$

❖ Représentation graphique

– Caractère qualitatif

Diagramme en barres



1.2 Statistique descriptive à une dimension

❖ Représentation graphique (suite)

– Caractère quantitatif discontinu

Diagramme en bâtons

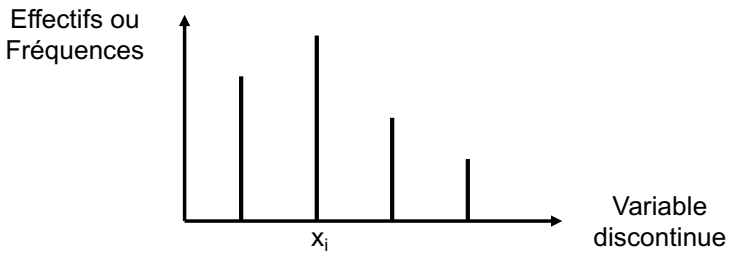
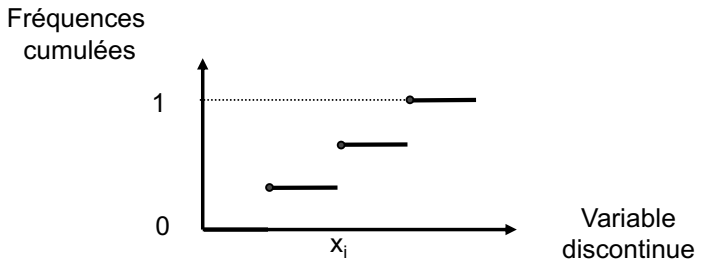


Diagramme des fréquences cumulées
(diagramme en escalier)



– Caractère quantitatif continu

Histogramme et polygone des fréquences

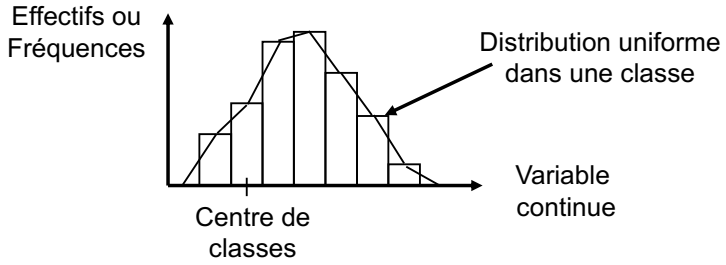


Diagramme des fréquences cumulées

1.2 Statistique descriptive à une dimension

❖ Caractéristiques de tendance centrale (de position)

– Moyenne arithmétique

$$\bar{x} = \frac{\sum_i n_i \, x_i}{\sum_i n_i}$$

– Médiane

valeur de la variable statistique qui partage la série de données (ordonnée en ordre croissant ou décroissant) en deux parties M_e

– Mode (ou classe modale)

Modalité ou valeur (ou classe) de la variable statistique la plus fréquente que l'on observe dans la série de données M_o

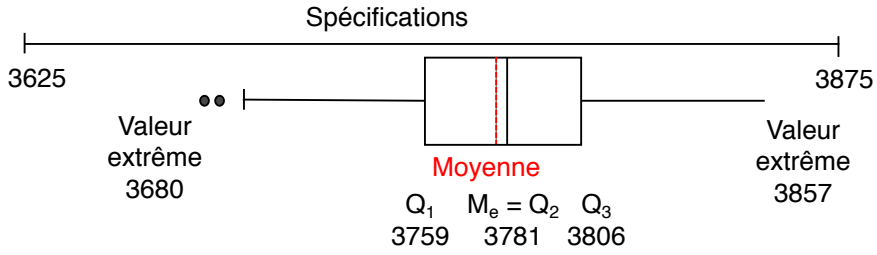
– Quantiles

valeurs de la variable statistique qui partagent la série de données (ordonnée en ordre croissant) en l parties égales

si l = 4	quartiles
si l = 10	déciles
si l = 100	centiles

❖ Le diagramme en boîte (Box-Plot)

spécification pression du gaz ambiant dans une lampe : 3750 mm Hg ± 125 mm
mesure de la pression sur un échantillon de 80 lampes



1.2 Statistique descriptive à une dimension

❖ Caractéristiques de dispersion

concentration des valeurs autour de la valeur centrale

– Etendue

$$E = x_{Max} - x_{min}$$

– Variance

moyenne des écarts quadratiques

$$V(X) = \sigma_X^2 = \frac{\sum_i n_i \left(x_i - \bar{x}\right)^2}{\sum_i n_i} = \frac{\sum_i n_i \, x_i^2}{\sum_i n_i} - \bar{x}^2$$

– Écart type

dispersion autour de la moyenne

$$\sigma_X = \sqrt{\frac{\sum_i n_i \left(x_i - \bar{x}\right)^2}{\sum_i n_i}}$$

– Coefficient de variation (en %)

homogénéité d'une distribution

$$CV = \frac{\sigma_X}{\bar{x}} . 100$$

❖ Caractéristiques de forme

– Distribution symétrique

moyenne = médiane = mode

– Coefficient d'asymétrie empirique de Pearson

$$S_k \cong \frac{3\left(\bar{x} - M_e\right)}{\sigma_X} \cong \frac{\bar{x} - M_o}{\sigma_X}$$

1.3 Statistique descriptive à deux dimensions

- *Traitement de deux variables :*
  répartition des mesures ou observations, synthèse et analyse
- *Données d'observations, expérimentales, séries statistiques ou résultats d'un sondage*

❖ Distribution statistique double

– Tableau de contingence

X / Y	y_j y_q	Loi marginale de X
x_i x_p	n_{ij}	$n_{i.}$
Loi marginale de Y	$n_{.j}$	$n_{..}$

– Moyennes marginales

$$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{p} n_{i.} \, x_i \qquad \bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{q} n_{.j} \, y_j$$

– Variances marginales

$$V(X) = \frac{1}{n_{..}} \sum_{i=1}^{p} n_{i.} \left(x_i - \bar{x}\right)^2 \qquad V(Y) = \frac{1}{n_{..}} \sum_{j=1}^{q} n_{.j} \left(y_j - \bar{y}\right)^2$$

– Moyennes conditionnelles

$$\bar{x}_j = \frac{1}{n_{.j}} \sum_{i=1}^{p} n_{ij} \, x_i \qquad \bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{q} n_{ij} \, y_j$$

– Variances conditionnelles

$$V_j(X) = \frac{1}{n_{.j}} \sum_{i=1}^{p} n_{ij} \left(x_i - \bar{x}_j\right)^2 \qquad V_i(Y) = \frac{1}{n_{i.}} \sum_{j=1}^{q} n_{ij} \left(y_j - \bar{y}_i\right)^2$$

1.3 Statistique descriptive à deux dimensions

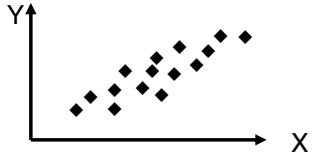
❖ Etude simultanée de deux dispersions

– Notion de corrélation

Il y a corrélation entre deux variables observées sur les éléments d'une même population lorsque les variations des deux variables se produisent **dans le même sens** (corrélation positive) ou lorsque les variations sont **de sens contraire** (corrélation négative)

– Diagramme de dispersion

- Forme de la liaison (linéaire)
- Intensité de la liaison



– Coefficient de corrélation

- Estimation du degré de corrélation linéaire entre deux variables aléatoires X et Y d'une même population

$$r = \frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_X . \sigma_Y}$$

avec

$$\mathrm{Cov}\left(X,Y\right) = \frac{1}{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}} \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\left(x_i - \bar{x}\right)\left(y_j - \bar{y}\right) = \frac{1}{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}} \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\, x_i.y_j - \bar{x}.\bar{y}$$

- Propriétés :
  - r est indépendant des unités de mesure de X et de Y
  - r = – 1 (corrélation négative parfaite)
  - r = + 1 (corrélation positive parfaite)
  - si deux variables sont indépendantes, r = 0
  - si r = 0, absence de corrélation **linéaire**

1.3 Statistique descriptive à deux dimensions

❖ Droite de régression linéaire – Méthode des moindres carrés

- Nuage de points : n couples (x_i ; y_i)
- Modèle construit sur les n données d'observations de

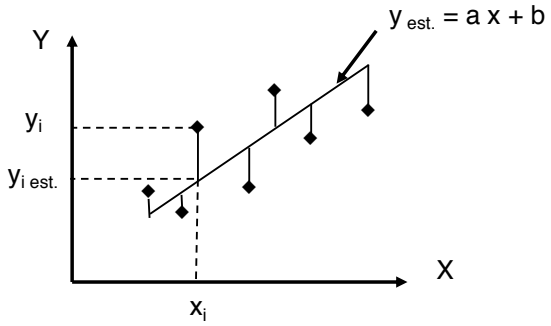
$$y_i = a\, x_i + b + e_i$$

- Droite de régression empirique

$$y_{est.} = a\, x + b$$

On cherche l'équation de la droite qui permet de rendre minimum la somme des carrés des écarts des valeurs observées y_i à la droite

Minimiser $\quad \sum_i e_i^2 = \sum_i \left(y_i - y_{i\,est.}\right)^2$



- Paramètres de la droite des moindres carrés :

$$a = \frac{\mathrm{Cov}\left(X,Y\right)}{V\left(X\right)} \qquad b = \bar{y} - a\,\bar{x}$$

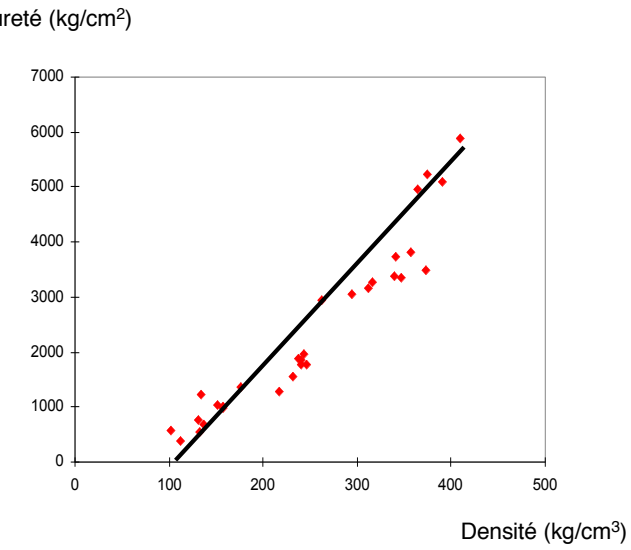
- On peut définir la droite : $x_{est.} = a'y + b'$

$$a' = \frac{\mathrm{Cov}\left(X,Y\right)}{V\left(Y\right)} \qquad b' = \bar{x} - a'\,\bar{y} \qquad a\,a' = r^2$$

1.3 Statistique descriptive à deux dimensions

Etude de propriétés mécaniques de panneaux de contre-plaqué

Obs n°	Densité (kg/m^3)	Dureté (kg/cm^2)
1	152	1041
2	134	1230
3	157	984
4	176	1366
5	133	532
6	158	997
7	137	683
8	102	567
9	112	372
10	131	754
11	294	3048
12	240	1780
13	243	1970
14	263	2938
15	357	3809
16	246	1779
17	240	1843
18	232	1557
19	237	1880
20	217	1268
21	410	5870
22	375	5223
23	390	5103
24	373	3481
25	312	3164
26	339	3389
27	365	4952
28	347	3350
29	317	3281
30	341	3729



$\bar{x} = 251$

$\bar{y} = 2398$

$r = 0{,}955$

$a = 15{,}61$

$b = -1521{,}41$

$y_{est.} = 15{,}61\ x - 1521{,}41$

Prévision de la dureté pour une densité de 240 kg/m^3 :

$y_{est.} = 2226{,}24\ kg/cm^2$