Statistiques

UNIVERSITÉ INTERNATIONALE DE CASABLANCA
U.I.C

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 et 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 a 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 mesuré.

✓Un échantillon est un sous-ensemble

- ✓de l'univers : s'il est composé d'unités expérimentales,
- ✓de la population : s'il est composé de mesures de la variable.

✓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 mesure 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 8 circuits électroniques de la production d'une usine et on a mesuré la longueur et la résistance a la traction des fils d'interconnexion de chaque circuit.

No. de l'observation	Resistance a 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 a 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 subdivisée 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
$a \leq x \leq b$				
....				

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.

✓ Mesure d'association : coefficient de corrélation.

Soit x1; x2; …..; xn un échantillon de n observations d'une population (valeurs numériques).

✓ La moyenne de l'échantillon, ou moyenne echantillonnale est:

$$(1/n)\sum_{i=1}^{n} xi$$

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

La médiane de l'échantillon, dénotée ~x, est une valeur telle que 50% des observations lui sont supérieures et 50% lui sont inferieures.

Si x(1); x(2); : : : ; x(n) sont les données en ordre croissant alors

~x = x((n+1)/2) si n est impair.

{x(n/2) + x((n/2) +1)}/2 si n est pair.

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.

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

Un échantillon peut avoir plusieurs modes.

✓Le mode est nécessairement égal a l'une des données.

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

Soit x1; x2; : : : ; xn un échantillon de n observations d'une population (valeurs numériques).

✓ L'étendue de l'échantillon est

$R = \max(x1; x2; : : : ; xn) - \min(x1; x2; : : : ; xn)$

✓L'ecart interquartile est

$$IQR = Q3 - Q1$$

ou Q1 et Q3 sont les premier et troisieme quartiles.

✓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 : Q2 = médiane de l'échantillon.

2. Poser

Q1 = médiane du sous-ensemble des valeurs inferieures a Q2.

Q3 = médiane du sous-ensemble des valeurs supérieures a Q2.

Dispersion : variance, écart-type, coefficient de variation

Soit x1; x2; : : : ; xn un échantillon de n observations d'une population (valeurs numériques).

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

$$S_{XX}=\sum_{i=1}^{n}(x_i-m)^2$$

$$s^2 =(1/n-1) \times S_{XX} =((1/n-1)\sum_{i=1}^{n}(x_i-m)^2)= (1/n-1)[(\sum_{i=1}^{n}x_i^2)-nm^2]$$

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 :

CV = s/m.

Moments (centres) : $\mu_k =(1/(n-1))\sum_{i=1}^{n}(x_i-m)^k$

μ_3 / s^3 (asymétrie)

μ_4 / s^4(aplatissement).

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 Etendue, moyenne , médiane, variance, les quartiles, IQR, le mode, et CV.

Association : coefficient de corrélation

Soit n observations de deux variables quantitatives (xi; yi) avec i = 1; 2; … ; n. Le coefficient de corrélation de x et y est

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\ \sqrt{S_{XX}}}$$

On pourra montrer que $-1 \leq r \leq 1$

Interprétation du coefficient de corrélation

✓Si $|r| = 1$ alors il y a corrélation parfaite entre les xi et les yi.

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 xi et les yi.

Les points du diagramme de dispersion sont distribues ‘’au hasard" dans le plan.

✓Si $-1 < r < 1$ alors il y corrélation forte, moyenne ou faible entre les xi et les yi.

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).