Descriptive généralites

Une variable statistique est dite :

Quantitative

- **quantitative** : lorsqu'elle est mesurée par un nombre (les Notes des Etudiants à l'Examen de Statistique, le Chiffre d'Affaire par PME, le Nombre d'Enfants par Ménage, . . .).

On distingue 2 types de variables quantitatives :

- les variables quantitatives discrètes
- Les variables quantitatives continues.

Par exemple le nombre d'enfants par ménage ne peut être que 0, ou 1, ou 2, ou 3, . . .C'est une variable quantitative discrète.

Les variables **quantitatives continues** peuvent prendre toute valeur dans un intervalle. Par exemple, le chiffre d'affaire par PME, le temps d'attente à un arrêt de bus.

Qualitative

Lorsque les modalités (ou les valeurs) qu'elle prend sont désignées par des noms. Par exemples, les modalités d'une variable Sexe peuvent être : Masculin et Féminin.

On distingue deux types de variables qualitatives : les variables qualitatives ordinales et les variables qualitatives nominales.

Plus précisément une variable qualitative est dite **ordinale**, lorsque ses modalités peuvent être classées dans un certain ordre naturel (c'est par exemple le cas d'une variable commen Mention au Bac).

Une variable qualitative est dite **nominale**, lorsque ses modalités ne peuvent être classées de façon naturelle (c'est par exemple le cas d'une variable comme Couleur des Yeux ou encore de la variable Sexe).

Exercice fréquence

Récupérez le dataset dans student.csv, dans le dossier data et étudiez le critère couleur des yeux.

1. Créez un DataFrame avec les données dans le dossier data.

2. Edutiez le critère “Couleur des yeux”. Créez un nouveau DataFrame dans lequel vous calculerez la fréquence d'apparition de ce critère.
3. Créez à partir du tableau précédent un diagramme en secteur pour visualiser la répartition de la couleur des yeux.

Exercice notes en statistique

Récupérez le dataset **note_statistiques.csv** et faite un diagramme en baton des notes en statistiques.

Renommez la colonne Notes examen de statistiques en Notes.

1. Que constatez-vous pour ce diagramme ?
2. Quelle est le type de la variable notes ?
3. Faites maintenant des classes de valeurs de largeur 4 pour les notes.
4. Créez à partir du dernier regroupement un historgramme.

La classe modale est la classe dont la fréquence par unité d'amplitude est la plus élevée.

Cette classe correspond donc au rectangle le plus haut de l'histogramme des fréquences.

Trouvez la classe modale de notre dataset notes statistiques.

Valeurs centrales

Définition du mode

Le mode correspond à la valeur de la variable pour laquelle l'effectif (ou la fréquence) est le plus grand.

Recensement des familles dans la population dont le nombre d'enfants de moins de 14 ans.

Nombre d'enfants	Nombre de famille
0	2601
1	6290
2	2521
3	849
4	137

Ici c'et la valeur 1.

Remarques

Certaines variables peuvent présenter plusieurs modes.

Exercice variable continue

Prix	Effectifs
[210, 230]	30
[210, 230]	60
[210, 230]	100
[210, 230]	20
	total = 210

Créez un histogramme à partir des données suivantes. Pensez à créer un DataFrame et définisser des classes à l'aide de la méthode cut de Pandas, par exemple. Et déterminez enfin le mode de la série statistique.

Quel est le type de la variable Prix ?

Médiane et Quantile

Ces objets sont des **indicateurs de position** en statistiques. Au même titre que la moyenne.

La médiane (notée Me) d'une variable quantitative est la valeur de cette variable qui permet de scinder la population étudiée en deux sous-populations de même effectif.

La notion de quantile d'ordre α ($0 \leq \alpha \leq 1$), encore appelée fractile d'ordre α, généralise la notion de médiane. Le quantile d'ordre α d'une variable quantitative X, est la valeur X_α de cette variable qui permet de scinder la population étudiée en deux sous-populations dont les effectifs respectifs sont égaux à α et $1 - \alpha$ de l'effectif de la population initiale.

Exercice Quartile Définition

En reprenant la définition d'un quantile ci-dessus essayez de décrire ce qu'est un quartile.

Correction

Les quartiles de X sont ses trois quantiles $x_{0,25}, x_{0,5} et x_{0,75}$. $Q_1 = x_{0,25}$, s'appelle le premier quartile; un quart des valeurs prises par X sont inférieures ou égales à Q_1 . $Q_2 = x_{0,5} = Me$ est la médiane. $Q_3 = x_{0,75}$ s'appelle le troisième quartile; un quart des valeurs prises par X sontsupérieures ou égales à Q_3.

Définition de l'intervalle interquartile

C'est la différence entre le troisième quartile et le premier quartile ; il s'écrit : $II = Q_3 - Q_1$.

Exercice Quartile notes

Soit la série suivantes : 10; 25; 30; 40; 41; 42; 50; 55; 70; 101; 110; 111

1. Déterminez le troisième quartile.
2. La médiane.
3. L'intervalle interquartile.

Même exercice avec la série statistique suivante, répondez aux 3 questions :

Soit la série suivantes : 10; 25; 30; 40; 41; 42; 50; 55; 70; 101; 110; 111, 208

Diagramme en moustache

La boîte à moustaches résume quelques **indicateurs de position** d'un caractère étudié (médiane, quartiles, minimum, maximum ou déciles).

Une boîte à moustache aide à se représenter les données. Notons que 50% des valeurs sont à l'intérieur de la boîte. De même 50% des valeurs sont positionnées sur les moustaches, en deux intervalles de 25%. De plus lorsqu'on réalise un diagramme à moustache, on voit directement les valeurs extrêmes.

Ce diagramme est utilisé **principalement** pour comparer un même caractère dans deux populations de tailles différentes.

Notez que les valeurs extèmes n'apportent en elle-même pas d'information particulière pour l'étude d'un critère.

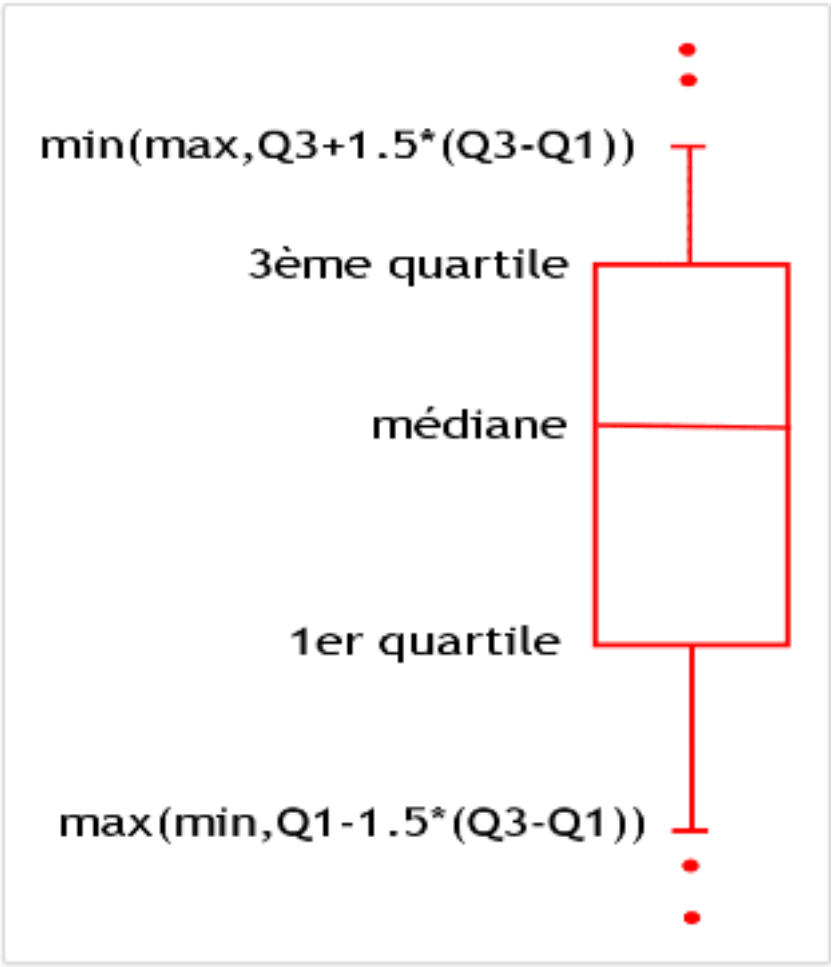

Figure 1: diagramme moustache

Exercice d'application

Utilisez seaborn. Rappelons que Seaborn est une bibliothèque pour faire des graphiques statistiques en Python. Il est construit au-dessus de matplotlib et étroitement intégré aux structures de données de Pandas.

```
%matplotlib notebook
import seaborn as sns
import matplotlib.pyplot as plt
# définiton d'un style de type grille sous fond blanc
sns.set(style="whitegrid")
```

Rappels de configuration pour vos graphiques dans Notebook :

- %matplotlib notebook conduira à des tracés interactifs intégrés au notebook.
- %matplotlib inline conduira à des images statiques de votre tracé dans un notebook.

Fixez également la taille de vos graphiques dans votre Notebook :

```
fig_dims = (20, 15)
fig, ax1 = plt.subplots(figsize=fig_dims)

ax2 = sns.boxplot(x="Name_variable1", y="Name_variable2", ax=ax1, data=MyDataFrame)
plt.show()
```

Utilisez la méthode boxplot de Seaborn :

```
# x et y sont des variables (colonnes) de votre DataFrame
# kind indique le type de graphique que vous souhaitez utiliser, ici box pour les diagrammes
# data représente le jeu de données
ax = sns.boxplot(x="Name_variable1", y="Name_variable2", kind="box", data=MyDataFrame)
plt.show()
```

Récupérez le jeu de données sur les tips intégré dans Seaborn à l'aide du code suivant :

```
tips = sns.load_dataset("tips")
```

1. Etudiez la position de la variable **total_bill** pour vous faire une idée du positionnement des valeurs.
2. Etudiez maintenant la position des variables du total des additions par jour.
3. Utilisez maintenant le paramètre **hue** dans la méthode boxplot. Il permet de faire un regroupement imbriqué par deux variables catégorielles. Etudiez le positionnement de la variable total_bill par jour, par rapport à la variable nominale (ou catégorielle pour Pandas et Seaborn) "smoker".

4. Récupérez maintenant le jeu de données suivant :

```
iris = sns.load_dataset("iris")
```

Etdudiez la position des variables suivantes : le spécimen et la taille des pétales.

Indicateurs de dispersion

On dispose d'une population de N individus, et on observe $x_1, ..., x_N$, les valeurs d'une variable quantitative discrète X pour ces individus.

Définition de l'étendue

L'étendue de la variable X quantitative discrète est la différence entre la plus grande et la plus petite des valeurs observées.

$e = max(X_i) - min(X_i)$ où i décrit l'ensemble des valeurs de la variable X.

Variance et écart type

La variance de la variable quantitative X, notée par Var(X), est, par définition, la moyenne arithmétique des carrés des écarts à la moyenne arithmétique.

$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $\bar{X}$ représente la moyenne.

Quand la série prend les valeurs $x_1, x_2, ..., x_k$ avec les fréquences $f_1, f_2, ..., f_k$, sa variance est :

$Var(X) = \sum_{i=1}^{k} f_i \times (X_i - \bar{X})^2$

Exercice calcul de la variance

1. Soit la série statistique suivante :

```
Notes = [7, 9, 11, 12, 13, 15]
```

Ecrire un script en Python qui permet de calculer la variance de cette série. Pensez à utiliser numpy.

2. Soit le tableau de notes d'une classe suivant :

Note	Effectif
7	5
9	4
11	21
12	35
13	32

Note	Effectif
15	3

2.1 Calculez la moyenne des notes de cette classe.

2.2 Déterminez la médiane. Faites un script en Python.

2.3 Déterminez la variance. Utilisez le code précédent en Python pour calculer la variance.

2.4 On définit l'écart type comme étant la racine de la variance. Calculez l'écart type des notes de cette classe.

3. L'enseignant augment toutes les notes de 1 point. Quelle conséquence cela a pour les calculs précédents ?