Chapitre 3
Introduction aux statistiques

1. Histoire et rôle de la statistique

1.1 Origines et évolutions

La statistique n'est pas une invention moderne. Son nom vient de la racine latine *status*, qui a d'abord donné en italien le mot *statista* (l'homme d'État), avant que l'allemand *Statistik* ne désigne, au XVIII^{e} siècle, la science descriptive des affaires de l'État. À l'origine, elle servait principalement à l'administration pour faire l'inventaire des populations ou des richesses. Ces dénombrements étaient cruciaux pour lever les impôts et les armées.

L'arrivée des ordinateurs dans la seconde moitié du siècle a changé l'échelle des calculs possibles. Des méthodes jusque-là impraticables deviennent accessibles : régression multiple, classification automatique, analyse factorielle. Les années 1980 et 1990 voient naître le machine learning, qui automatise la construction de modèles prédictifs à partir de grandes quantités de données.

Aujourd'hui, la statistique irrigue presque tous les secteurs. La santé publique l'utilise pour évaluer l'efficacité des traitements. Les entreprises l'appliquent pour segmenter leurs clients ou optimiser leur logistique. Les sciences sociales l'emploient pour mesurer l'impact de politiques publiques.

Python s'inscrit dans cette continuité historique. Il donne accès aux méthodes développées au cours du XX^{e} siècle, tout en simplifiant leur mise en œuvre. Les bibliothèques que nous utiliserons encapsulent des décennies de recherche mathématique. Elles permettent d'appliquer des tests, de construire des modèles, de visualiser des résultats sans réécrire les algorithmes sous-jacents.

1.2 Statistique descriptive et statistique inférentielle

La statistique se divise en deux approches complémentaires. Chacune répond à des questions différentes et mobilise des outils distincts.

1.2.1 Statistique descriptive

La statistique descriptive résume les données disponibles. Elle ne cherche pas à généraliser au-delà de ce qui est observé. Son objectif : rendre l'information lisible.

Prenez un fichier contenant les ventes mensuelles d'une entreprise sur trois ans. La statistique descriptive calcule la moyenne des ventes, identifie le mois le plus performant, mesure la dispersion autour de la moyenne. Elle produit des tableaux, des graphiques, des indicateurs chiffrés. Ces éléments aident à comprendre ce qui s'est passé.

Les outils principaux sont simples mais puissants. La moyenne donne une valeur centrale. La médiane résiste mieux aux valeurs extrêmes. L'écart-type mesure la dispersion. Les quantiles découpent les données en segments égaux. Les histogrammes montrent la distribution. Les diagrammes en boîte révèlent les valeurs atypiques.

Cette approche ne requiert aucune hypothèse probabiliste. Elle décrit des faits. Si vous avez mesuré la taille de 200 personnes, la statistique descriptive indique que la taille moyenne est de 1,72 mètre dans ce groupe de 200 personnes. Elle ne prétend rien sur les personnes non mesurées.

La statistique descriptive excelle dans trois situations. D'abord, lors de l'exploration initiale des données. Avant toute analyse complexe, il faut comprendre ce qu'on manipule : quelles variables, quelle échelle, quelles valeurs manquantes, quels outliers. Ensuite, pour la communication de résultats. Un graphique bien construit transmet plus d'information qu'un paragraphe. Enfin, quand la population entière est accessible. Si vous analysez tous les employés d'une entreprise, pas besoin d'inférence : vous avez la totalité des données.

Si nous reprenons les données de la sous-section précédente, nous avons un chiffre d'affaires par date et par boutique.

date	shop	visitors	revenue
2023-01-02	Paris 1	170	1 366,09 €
2023-01-02	Paris 2	243	1 329,91 €
2023-01-02	Paris 3	179	2 577,22 €
2023-01-02	Lyon	273	4 144,97 €
2023-01-03	Lyon	308	5 145,73 €
2023-01-04	Paris 1	202	2 738,64 €
2023-01-04	Paris 2	194	848,54 €
2023-01-04	Paris 3	199	940,67 €
2023-01-05	Paris 1	166	1 773,88 €
2023-01-05	Paris 2	189	1 187,66 €
2023-01-05	Paris 3	196	1 506,38 €

Imaginons que ce tableau prend l'ensemble des données disponibles. On peut remarquer que le chiffre d'affaires maximum est de 5 145,73 € le 3 janvier 2023 à la boutique de Lyon et que le chiffre d'affaires minimum est de 848,54 € le 4 janvier 2023 à Paris 2.

1.2.2 Statistique inférentielle

La statistique inférentielle va plus loin. Elle tire des conclusions sur une population entière à partir d'un échantillon. Elle pose une question : que peut-on dire de ce qu'on n'a pas observé ?

Reprenons l'exemple des tailles. Vous mesurez 200 personnes tirées au hasard dans une ville de 100000 habitants. La moyenne observée est 1,72 mètre. La statistique inférentielle estime la moyenne de la population complète et fournit un intervalle de confiance : « La taille moyenne se situe probablement entre 1,70 et 1,74 mètre. » Elle quantifie l'incertitude liée à l'échantillonnage.

Cette approche repose sur la théorie des probabilités. Elle suppose que l'échantillon est représentatif, c'est-à-dire tiré aléatoirement. Elle modélise la variabilité naturelle des données. Elle calcule la probabilité d'observer certains résultats si telle hypothèse est vraie.

Les applications sont multiples. Les sondages d'opinion estiment les intentions de vote en interrogeant quelques milliers de personnes. Les essais cliniques évaluent l'efficacité d'un médicament sur quelques centaines de patients, puis généralisent aux millions de malades potentiels. Les tests qualité dans l'industrie vérifient qu'une chaîne de production fonctionne correctement en contrôlant une fraction des pièces produites.

La statistique inférentielle mobilise plusieurs techniques. Les tests d'hypothèses vérifient si un effet observé est significatif ou dû au hasard. Les intervalles de confiance encadrent une valeur inconnue avec un degré de certitude. Les modèles de régression prédisent une variable à partir d'autres. L'ANOVA compare les moyennes de plusieurs groupes.

Remarque

Ces deux branches ne s'opposent pas, elles se complètent. Toute analyse commence par une phase descriptive. On examine les données, on détecte les anomalies, on visualise les distributions. Cette étape oriente la suite : quel test appliquer, quel modèle construire, quelles transformations effectuer. L'inférentielle prend ensuite le relais quand on veut généraliser. Mais elle s'appuie sur les résultats descriptifs. Un intervalle de confiance utilise la moyenne et l'écart-type calculés sur l'échantillon. Python traite ces deux branches différemment. La statistique descriptive mobilise principalement Pandas et NumPy : calcul de moyennes, de quantiles, de corrélations. La statistique inférentielle fait appel à SciPy Stats et statsmodels : tests, intervalles de confiance, modèles. Matplotlib et seaborn servent les deux en produisant les visualisations.

2. Nature des données

Avant toute analyse, identifier le type de données manipulées détermine les méthodes applicables. Une moyenne sur des codes postaux n'a aucun sens. Un histogramme sur des catégories non ordonnées brouille la lecture. Cette section clarifie les distinctions essentielles. Les données se répartissent en deux grandes familles selon leur nature : qualitatives ou quantitatives. Cette distinction oriente directement le choix des outils statistiques.

2.1 Variables qualitatives

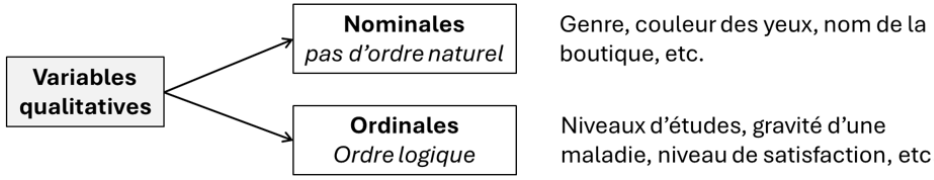
Une **variable qualitative** décrit une caractéristique non numérique. Elle classe les observations en catégories distinctes. Ces catégories peuvent être des mots, des codes, des symboles. Par exemple, on retrouve dans cette catégorie : le genre, la couleur des yeux, le type de contrat, la région géographique ou le niveau de satisfaction.

Les variables qualitatives se subdivisent en deux sous-catégories. Les **variables nominales** n'ont pas d'ordre naturel entre les catégories. Le genre, la couleur des yeux ou le type de contrat sont nominaux. Aucune catégorie n'est supérieure ou inférieure à une autre. On peut seulement compter les effectifs dans chaque catégorie.

Les **variables ordinales** possèdent un ordre logique. Le niveau de satisfaction, le niveau d'études (primaire, secondaire, supérieur), la gravité d'une maladie (légère, modérée, sévère) sont ordinaux. On sait qu'une catégorie est « plus » ou « moins » qu'une autre, mais l'écart entre catégories n'est pas mesurable. La différence entre « insatisfait » et « neutre » n'équivaut pas nécessairement à celle entre « neutre » et « satisfait ».

Les opérations possibles sur ces variables restent limitées. On calcule des fréquences, des pourcentages, des modes (la catégorie la plus fréquente). On peut construire des tableaux croisés pour étudier les relations entre deux variables qualitatives. Les graphiques appropriés sont les diagrammes en barres ou les diagrammes circulaires.

Calculer une moyenne n'a pas de sens. Si vous codez les genres par des chiffres (1 pour homme, 2 pour femme), la moyenne de ce codage ne signifie rien. Ce piège est fréquent quand on transforme des catégories en nombres pour les traiter informatiquement.

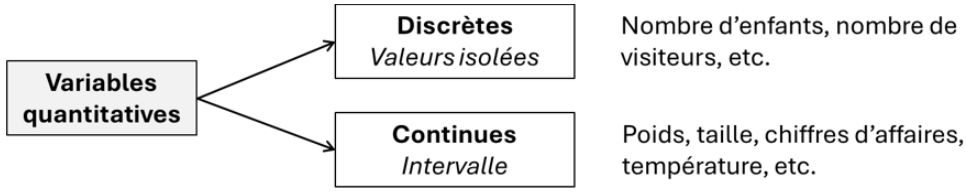


2.2 Variables quantitatives

Une **variable quantitative** mesure une quantité. Elle prend des valeurs numériques sur lesquelles les opérations arithmétiques ont un sens. On peut additionner, soustraire, calculer des moyennes ou des écarts-types. Par exemple, on retrouve dans cette catégorie : l'âge, le revenu, le poids, la température, le temps de réponse, le nombre de visites sur un site web, le chiffre d'affaires mensuel.

Les variables quantitatives se divisent aussi en deux types. Les **variables discrètes** prennent des valeurs isolées, généralement des entiers. Le nombre d'enfants dans un foyer, le nombre de visiteurs par jour, le nombre de pièces défectueuses dans un lot. On ne peut pas avoir 2,3 enfants ou 15,7 clients.

Les **variables continues** peuvent prendre n'importe quelle valeur dans un intervalle. Le poids, la taille, le temps, la température, le chiffre d'affaires sont continus. Entre deux valeurs, il existe toujours une valeur intermédiaire. En pratique, les instruments de mesure limitent la précision (on pèse au gramme près, on mesure au millimètre près), mais la variable reste conceptuellement continue.



Cas ambigus

Certaines variables posent question. Un code postal est numérique mais qualitatif : on ne peut pas calculer la moyenne des codes postaux d'un quartier. Une échelle de Likert (notation de 1 à 5 pour mesurer un accord) est parfois traitée comme quantitative par commodité, bien qu'elle soit strictement ordinale. Cette approximation est acceptable si l'échelle comporte suffisamment de niveaux et si les répondants la perçoivent comme régulièrement espacée.

L'âge illustre une transformation courante. Mesuré en années, c'est une variable quantitative discrète. Regroupé en tranches (0-18, 19-35, 36-60, 60+), il devient qualitatif ordinal. Cette transformation simplifie certaines analyses mais perd de l'information. Le choix dépend de l'objectif : conserver la précision ou faciliter l'interprétation.

Remarque

Dans notre exemple du chiffre d'affaires, la date peut être considérée comme une variable qualitative ordinale, le nom de la boutique comme une variable qualitative nominale, le nombre de visiteurs comme une variable quantitative discrète et le chiffre d'affaires comme une variable quantitative continue.

3. Statistique descriptive

Les **statistiques descriptives** transforment des données brutes en informations synthétiques. Cette section présente les méthodes essentielles pour analyser une variable isolément, puis pour étudier les relations entre deux variables.

3.1 Analyse univariée

L'**analyse univariée** examine une seule variable à la fois. Elle résume sa distribution, identifie ses caractéristiques centrales et mesure sa dispersion. Ce sont les premiers calculs à effectuer lorsque l'on importe un jeu de données.

3.1.1 Mesures de tendance centrale

Les **mesures de tendance centrale** résument une distribution par une valeur typique. Trois indicateurs dominent : la moyenne, la médiane et le mode.

Moyenne arithmétique

La **moyenne arithmétique** additionne toutes les valeurs et divise par le nombre d'observations. Elle représente le centre de gravité des données. Mathématiquement : si on note $x_1, x_2, \ldots, x_n$ les n valeurs, la moyenne vaut :

$$\text{moyenne} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Prenons l'exemple d'un taux de bonnes réponses à un examen sur l'ensemble des cinq apprenants de la session : 45%, 55%, 48%, 58%, 55%. Pour calculer la moyenne, nous effectuons le calcul suivant :

$$\text{moyenne} = \frac{45\% + 55\% + 48\% + 58\% + 55\%}{5} = 52{,}2\%$$

Attention, la moyenne est sensible aux valeurs extrêmes. Un taux très élevé tire la moyenne vers le haut, même si la majorité des taux reste autour de 50%.

Par exemple, imaginons que le meilleur taux n'est pas 58% mais 95%, alors la moyenne augmente significativement, passant de 52,2% à 59,6% :

$$\text{moyenne} = \frac{45\% + 55\% + 48\% + \mathbf{95}\% + 55\%}{5} = 59{,}6\%$$

La moyenne est alors fortement influencée par cette valeur extrême.

Bonne nouvelle, dans python, il n'est pas nécessaire de coder cette fonction. En effet, la moyenne est native dans NumPy ou Pandas.

Reprenons notre première série de taux de bonnes réponses, on peut utiliser la méthode `np.mean` pour calculer la moyenne.

```
import numpy as np
success_rates = np.array([0.45, 0.55, 0.48, 0.58, 0.55])
print(np.mean(success_rates)) # 0.522
```

Avec Pandas, la méthode s'applique directement sur une colonne. Reprenons l'ensemble des données de chiffre d'affaires du contenu dans le fichier sales_global.csv et calculons le chiffre d'affaires moyen avec la méthode `.mean()` :

```
import pandas as pd
sales = pd.read_csv("sales_global.csv", sep=';')
print(sales["revenue"].mean()) # 2719.600092339261
```

Médiane

La **médiane** divise les observations en deux groupes égaux. La moitié des valeurs lui est inférieure, l'autre moitié supérieure. Pour la calculer, on ordonne les données. Si le nombre d'observations est impair, la médiane est la valeur centrale. Si pair, c'est la moyenne des deux valeurs centrales.