LES STATISTIQUES

Les statistiques sont comme une arme de gros calibre : utile quand elles sont utilisées de façon appropriée mais dangereuses dans de mauvaises mains.

Si on veut comprendre les nombres derrière les nouvelles et apprécier le pouvoir des données, les statistiques sont incontournables.

Il est facile de mentir avec des statistiques mais il est difficile de dire la vérité sans elles.

Index : outil qui permet de condenser plusieurs informations en un seul nombre. C'est un outil de comparaison.

Les statistiques permettent de gérer des données, de donner de l'information.

Les données de plus en plus nombreuses aujourd'hui sont la matière première de la connaissance. Les statistiques sont l'outil le plus puissant pour donner du sens aux informations reçues :

- Description et comparaison : la stat descriptive permet la simplification qui cependant implique une certaine perte de nuance ou de détail.
- Déduction: une des fonctions clé de la stat est d'utiliser les données que nous avons afin de faire des conjectures sur des questions plus larges pour lesquelles nous ne disposons pas de toute l'information nécessaire. On utilise les données d'un monde « connu » afin d'en déduire un monde « inconnu ». Pour ce faire, l'échantillonnage sera utilisé = rassembler des données sur une petite zone et d'en extraire une généralité.
- Définir le risque et autre événements probables : le fondement en sera les probabilités, un outil qui nécessite un jugement adéquat.
- Identifier les relations importantes : comparable à un travail de détective, les données présentent des clés et des schémas qui permettent d'en tirer des conclusions censées. L'analyse de la régression sera l'outil qui permettra d'isoler des relations entre 2 variables mais ne

permettra pas d'expliquer pourquoi cette relation existe et s'il existe un lien de cause à effet entre elles.

Les statistiques nous donnent l'opportunité d'obtenir des infos compréhensibles sur notre vie.

En résumé, les statistiques servent à :

- Résumer une grande quantité de données,
- Prendre de meilleures décisions,
- Répondre à des questions d'ordre social,
- Reconnaître des schémas récurrents,
- Evaluer l'efficacité de politiques, programmes.

I) La statistique Descriptive

Le principe repose sur des nombres et des calculs afin de résumer des données brutes.

La première tâche est souvent de trouver une tendance centrale dans une distribution (série de données) = la Moyenne. Mais celle-ci est sujette à distorsion dans le cas de données anormales, trop éloignées. On lui préfère donc la Médiane, qui est le point divisant une distribution en 2 moitiés égales d'observations.

Dans le cas où aucune donnée anormale (excentrée) n'est présente, la moyenne et la médiane seront très proches. Les calculs sont simples dans les 2 cas mais on choisira l'une plutôt que l'autre mesure en fonction de la situation. L'avantage de la médiane est l'utilisation de sous-section telle que les déciles, quartiles…

Un nombre absolu est interprété sans contexte particulier ou autre information.

Un nombre relatif nécessite une comparaison ;

La déviation standard mesure la dispersion de données par rapport à la moyenne. Elle permet d'assigner un nombre à cette dispersion.

La Distribution représente la dispersion des données. La plus commune est la distribution normale avec des données symétriques autour de leur moyenne en forme de cloche.

Les statistiques descriptives sont souvent utilisées pour comparer deux chiffres ou des quantités. Si le contexte est absent, l'utilisation des pourcentages sera préférable. Mesurer un changement en pourcentage donne une échelle.

Un Index est une stat descriptive composée d'autres statistiques descriptives. Il permet la comparaison, à travers un simple nombre qui agrège de nombreuses informations complexes. Il sera sensible aux données qui le composent et au poids qui leur sera attribué.

La stat descriptive permet de mettre un cadre simplifié autour de données éparses et complexes.

Mais bien que le champ des statistiques soit ancré dans les mathématiques et que ces dernières soient exactes, l'utilisation des statistiques pour décrire des phénomènes complexes n'est pas exacte. Il reste beaucoup de marge pour voiler la vérité.

L'exactitude détermine si une information est cohérente avec la vérité. Mais la précision n'est pas gage de justesse. En fait, la précision peut masquer l'exactitude des faits en donnant une fausse impression de certitude, sciemment ou non.

Ce qui signifie que le plus précis des calculs doit être vérifié à l'aune du sens commun. Un des problèmes fondamental en ce qui concerne la statistique descriptive est le manque de clarté sur ce que l'on cherche à définir, décrire ou expliquer> cela implique de faire attention à l'unité d'analyse.

La Médiane est le point moyen de la distribution : 50% des observations seront au-dessus et 50% en dessous. Elle ne sera donc pas sensible aux éléments anormaux comme peut l'être la Moyenne.

L'emploi de l'une ou l'autre sera déterminé si on souhaite faire apparaître l'influence d'éléments anormaux dans le résultat.

De même, les chiffres utilisés, notamment pour des données monétaires, doivent être précis. Ainsi un chiffre nominal ne prend pas en compte l'inflation contrairement au chiffre réel.

Les pourcentages sont préférés aux données brutes car ils expriment un changement de quantité par rapport à une échelle, un contexte.

Mais comme dans toute notion de changement, il est nécessaire de bien définir le point de référence de départ ainsi que celui de fin de l'observation.

Un index statistique possède une distorsion introduite par la combinaison de plusieurs indicateurs en un nombre unique. Tout index est sensible à sa construction. Il sera affecté à la fois par les données qui entrent dans son calcul mais également par le poids de celles-ci.

La Corrélation mesure le degré qui relie deux phénomènes. Deux variables seront positivement corrélées si un changement dans l'une entraîne le changement de l'autre dans la même direction. Une corrélation sera négative si le changement de l'une des variables induit un changement de direction contraire dans l'autre variable. Elle s'exprimera par un nombre simple : le coefficient de corrélation compris entre – 1et +1.

Un coefficient de -1 est une parfaite corrélation négative, et +1 pour une parfaite corrélation positive.

0 signifie que les 2 variables n'ont aucune relation entre elles.

L'intérêt de ce coefficient est qu'il est indépendant de toute unité.

Mais une corrélation n'implique pas une causalité entre les variables.

II) Les Probabilités

C'est l'étude d'événements et résultats incluant un élément d'incertitude.

Elles ne nous disent pas ce qu'il va se passer de façon certaine, mais ce qu'il est susceptible de se passer ou ce qui est susceptible de moins se passer.

C'est le rapport entre l'occurrence d'un événement indépendant sur la totalité des possibilités. (Sortir un deux avec un dé est de 1/6).

La probabilité que 2 événements indépendants se produisent est le produit de leur probabilité respective.

La probabilité qu'un événement OU un autre se produise est la somme de leur probabilité respective.

Les probabilités permettent également de calculer l'outil le plus utile dans l'aide à la décision : la valeur espérée (l'espérance) qui est la somme des différents résultats, chacun pondéré par sa propre probabilité.

Un important théorème connu sous le nom de « loi des grands nombres » stipule que le plus le nombre d'essais augmente, plus la moyenne des résultats ce rapproche de l'espérance.

Les probabilités sont donc un outil pour gérer les incertitudes de la vie. Elles ne sont pas déterministes.

Les statistiques ne peuvent être plus intelligentes que les personnes qui les utilisent. Un exemple récent nous a été donné par la crise financière de 2008 avec un baromètre du risque qui a été largement utilisé : le Value at Risk (VaR), qui combinait un indicateur (plusieurs infos traduites en un seul nombre) avec la puissance des probabilités (attachant une espérance de gain/perte).

Sur la base de données des mouvements passés des marchés, les experts donnaient un chiffre en Dollar représentant ce que l'institution pouvait perdre sur une position sur une période de temps déterminée avec une probabilité de 99%. Le modèle VaR quantifiait le risque global.

Le problème est que le risque sous-jacent associé aux marchés financiers n'est pas aussi prévisibles qu'un lancer de pièces. La fausse précision du modèle créa un sentiment de fausse sécurité.

D'abord, les probabilités sous-jacentes sur lesquelles étaient bâtis les modèles, étaient basées sur les mouvements passés des marchés. Or il est connu, que sur ce type de marchés, le passé n'est pas gage du futur.

Les plus grands risques ne sont jamais ce que l'ont peut voir ou mesurer, mais ce que l'on ne voit pas et donc que l'on ne peut mesurer. Ceux qui se trouvent en dehors de toute frontière des probabilités normales peuvent se produire.

Les quants de WSt firent trois erreurs fondamentales : ils ont confondu exactitude et précision ; les estimations des probabilités sous-jacentes étaient erronées ; la « queue » de risque a été sous-estimée.

Ce ne sont pas les probabilités qui font des erreurs mais ceux qui les emploient.

Les principales erreurs d'interprétation sont :

- Assumer que les événements sont indépendants les uns des autres.
- Ne pas comprendre quand les événements sont effectivement indépendants. La définition statistique de l'indépendance entre 2 événements est le résultat de l'un ne doit pas influencer le résultat de l'autre. Il faut savoir faire la différence entre la perception et la réalité empirique.
- Des zones de confluence peuvent se produire : quand on assiste à un événement tout à fait hors contexte, on suppose qu'autre chose que le hasard est responsable.
- Le contexte entourant la statistique est négligé.
- Retour à la moyenne : les probabilités nous enseignent que toute observation très éloignée de la moyenne a de forte chance d'être suivie d'observations plus en rapport avec la moyenne de LT.
- La discrimination statistique : il ne faut pas oublier l'implication sociale des statistiques.

Le point à retenir est que notre habilité à analyser des données est devenue beaucoup plus sophistiquée que ce que nous devons faire des résultats. Nous aimons voir les nombres comme des faits froids et bruts. Si nos calculs sont justes, alors nous devons avoir la bonne réponse. Ce n'est pas toujours le cas. Il faut garder à l'esprit quel genre de calculs nous faisons et pourquoi.

D'où l'importance des données qui seront utilisées dans les calculs.

La base des études statistiques se font sur échantillonnage d'une population ciblée. Celui-ci se doit d'être le plus représentatif possible. Pour ce faire, il sera choisi de façon aléatoire, la clé de cette méthodologie étant que chaque observation d'une population concernée doit avoir une chance égale d'être incluse dans l'échantillon. Ce dernier doit ressembler le plus possible dans sa composition à la population de laquelle il est issu.

La quantité joue également un rôle et le plus grand sera l'échantillon, le meilleur il sera car les données hors normes seront lissées dans la totalité.

Les données utilisées doivent être source de comparaison : on choisira alors 2 groupes dont l'un sera le témoin, ceci afin d'isoler l'impact d'un point particulier. Et ces groupes seront définis de façon aléatoire.

Deux types d'études existent :

- Une étude longitudinale qui collecte des informations sur un large effectif à différentes périodes dans le temps. Ces données sont particulièrement intéressantes quand il s'agit d'étudier des relations de cause à effet, ce qui peut prendre parfois plusieurs années pour apparaître.
- Des données croisées qui sont des données collectées en un seul point du temps.

Derrière chaque étude importante se trouvent de bonnes données qui rendent les analyses possibles. Ce ne sont pas les statistiques qui « mentent » mais ce sont les données qui peuvent ne pas être adaptées à l'étude.

En fait il existe un biais de sélection où l'échantillon choisi n'appartiendra pas à la bonne population ou sera biaisé par les caractéristiques de celle population.

Ensuite, il existe un biais dans les publications des études où bien souvent ne seront publiées que les informations les plus attendues mais non pas la totalité. De fait, celles qui iront à l'encontre des premières ne seront pas connues.

La mémoire peut également faire défaut, ce pour quoi sont souvent préférées les études longitudinales car plus étendue dans le temps.

En conclusion, si la statistique est un travail de détective, les données en sont la clé.

III) Le Théorème Central Limite

Le cœur du principe de ce théorème est qu'un large et correct échantillonnage ressemble à la population générale dont il est issu. Des variations peuvent apparaître d'un échantillon à l'autre, mais la probabilité que l'un d'entre eux s'éloigne trop de celle-ci est très faible.

Si l'on possède des informations sur une population, il est possible d'en déduire des conclusions sur un échantillon extrait de cette même population.

Et réciproquement, des informations sur une population générale permettent de faire des déductions sur un échantillon issu de cette même population.

Les résultats tirés d'un échantillon seront donc un bon outil de déduction sur la population générale dont il extrait, les variations étant faibles.

Si nous avons des données sur une population et un échantillon, il est possible de savoir si ce dernier est extrait de cette population.

De même, si nous avons des données sur 2 échantillons, il est possible de savoir s'ils appartiennent à la même population.

Suivant ce théorème, la moyenne d'un échantillon sera distribuée globalement selon la loi normale autour de la moyenne de la population.

Et plus les échantillons sont importants en taille et en nombre, plus ils seront proches de la distribution normale.

Pour mesurer la dispersion de la moyenne des échantillons, on calcule « l'erreur standard » : alors que la déviation standard mesure la dispersion d'une population, l'erreur standard mesure la dispersion de la moyenne de

l'échantillon. Cette dernière est la déviation standard de la moyenne de l'échantillon.

L'erreur standard sera élevée quand la déviation standard de la distribution sous-jacente sera élevée. Un échantillon important issu d'une population très dispersée, sera aussi dispersé.

IV) Déductions

Elles nous disent ce qui est possible et ce qui ne l'est pas. C'est le procédé par lequel les statistiques nous parlent afin de nous donner des conclusions.

La déduction est l'assemblage de 2 concepts clés : les données et les probabilités, avec l'aide du théorème central limite.

Un des outils le plus utilisé est le test d'hypothèse : les statistiques ne peuvent pas prouver un fait à elles seules. On utilise donc les déductions statistiques afin d'accepter ou de rejeter des explications sur la base de leur probables véracités. On parlera ainsi d'hypothèse nulle = rejet d'une explication ou hypothèse alternative, qui sont complémentaires.

Afin de rejeter une hypothèse nulle, on défini un niveau significatif de 5% qui représente la probabilité d'observer un certain nombre de données si l'hypothèse nulle est vraie.

La p-value calculée sur le résultat est la probabilité spécifique d'obtenir un résultat aussi extrême que celui qui aurait été observé si l'hypothèse nulle était vraie.

Si on peut rejeter une hypothèse nulle avec un degré de signification suffisant, le résultat sera considéré comme statistiquement significatif.

Cependant, la signification statistique ne donne pas d'information quant à la taille de l'association> dire qu'il n'y a pas d'association statistiquement significative entre 2 variables implique que cette relation n'est due qu'à la chance.

La déduction statistique est un formidable outil qui permet de donner sens au monde en déterminant les explications les plus plausibles entre différentes variables.

V) Les Sondages

Un sondage est une déduction de l'opinion d'une partie de la population basée sur les réponses exprimées par un échantillon tiré de cette population appuyé par le Théorème Central Limite.

La différence entre un sondage et toute autre forme d'échantillonnage est que l'échantillon statistique ne s'exprimera pas sous forme de moyenne mais de pourcentage. Sinon, le procédé d'analyse sera identique.

De plus, plus l'échantillon est important, plus l'erreur standard sera réduite, ce qui donnera des résultats d'autant plus précis, proche de la réalité.

De mauvais résultats ne proviendront pas de mauvais calculs sur les erreurs standards mais sur l'utilisation de mauvais échantillons.

Les questions méthodiques à se poser seront donc :

- Est-ce que l'échantillon est bien représentatif de la population dont on essaie de mesurer l'opinion ? > il faudra faire attention à l'exclusion de certains segments de la population, valider la quantité de réponses.
- Est-ce que les questions ont été posées de telle façon à être le plus précises possibles quant à l'information donnée ? >la façon de formuler la question et le vocabulaire employé auront une influence sur la compréhension de la question et l'orientation de celle-ci.
- Les répondants disent-ils la vérité ?

La véritable difficulté des sondages est de construire un échantillon représentatif et d'en extraire l'information parfaitement représentative de la population.

Les résultats devront toujours être analysés dans leur contexte.

VI) La Régression

Ce type d'analyse permet de quantifier les relations entre des variables particulières et un résultat que l'on recherche en contrôlant les autres facteurs pouvant influencer ces derniers : on essaiera d'isoler l'effet d'une variable en gardant les autres constantes.

Cela ressemble un peu à la technique de sondage avec un échantillon large représentatif et une méthodologie solide qui évite des déviations trop importantes par rapport à la population générale.

La régression permet donc d'isoler les relations statistiques que l'on recherche en établissant la meilleure relation linéaire entre les différentes variables. Pour ce faire, la méthode des moindres carrés sera appliquée qui minimise la somme des carrés des résidus (distance entre le point et la droite de régression doit être la plus faible possible)

On distingue deux sortes de variables :

- La variable *dépendante* : celle qui est expliquée car elle dépend de différents facteurs ;
- La variable *explicative* : donne l'explication de la variable dépendante.

Afin de relier ces deux types de variables, on utilise un coefficient de régression qui dit que l'augmentation d'une unité de la variable explicative sera associée à l'augmentation de x unités de la variable dépendante. On s'intéressera à 3 critères :

- Le signe du coefficient qui donnera le sens de l'association entre les variables.
- La taille : quelle est l'importance de l'effet observé entre les variables dépendantes et explicatives (indépendantes).
- La signification : est-ce que la relation décrite a un sens qui est susceptible d'être observée dans la population générale.

On pourra donc être amené à calculer une erreur standard du coefficient de régression qui sera la mesure de la dispersion probable du coefficient si cette analyse de régression était répétée sur différents échantillons i.e. que 95% des

coefficients de régression observés se trouveront dans l'intervalle des erreurs standard de 2.

Un coefficient de régression sera d'autant plus significatif qu'il aura au moins deux fois la taille de l'erreur standard.

Une fois que le coefficient et l'erreur standard sont établis, l'hypothèse nulle peut être testée, selon laquelle aucune relation n'existe entre la variable dépendante et la variable explicative.

Un autre type de mesure sera employée avec l'analyse de régression : R^2 qui exprime le montant total de la variation expliquée par l'équation de régression, de combien est cette variation autour de la moyenne.

L'analyse en régression permet d'expliciter des relations complexes dans les quelles des facteurs multiples affectent les résultats voulus. Quand plusieurs variables explicatives sont impliquées, on parlera d'analyse en régression multiple qui donnera un coefficient de régression pour chacune des variables incluse dans l'équation de régression. A chaque ajout d'une variable explicative, les coefficients de corrélation seront calculés afin de minimiser la somme des moindres carrés de l'équation.

L'analyse de régression domine les méthodes scientifiques. C'est l'outil le plus important pour trouver des configurations sensées à un grand nombre de données.

Parce que l'analyse de régression est la bombe de l'arsenal statistique et qu'elle donne des réponses précises mais plus ou moins exactes à des questions complexent et afin qu'elle garde tout son sens, plusieurs erreurs sont à éviter :

- utiliser la régression pour analyser des relations non linéaires : le coefficient de régression mesure la pente de la droite de régression. Il ne peut donc y avoir plusieurs pentes et donc les données doivent être linéaires.

- une corrélation n'induit pas une causalité : l'analyse en régression ne décrit que l'état d'une relation entre deux variables.

- une causalité inverse : une relation entre A et B ne prouve pas que B est la cause de > les variables explicatives utilisées ne doivent pas être affectées par

le résultat de ce que l'on cherche. Ce sont les variables explicatives qui affectent la variable dépendante et non l'inverse.

- Omettre les biais > les résultats de la régression seront trompeurs et inexacts si l'équation de régression laissent de côté des variables explicatives importantes, d'autant si d'autres variables de l'équation prennent compte de cette omission.

- Variables explicatives hautement corrélées : l'analyse ne pourra pas alors distinguer la véritable relation entre ces variables et le résultat recherché.

- Extrapoler au-delà des données : les résultats ne seront valides que dans un environnement équivalent à l'échantillon sur lequel l'analyse a été pratiquée.

- L'utilisation d'un trop grand nombre de variables avec le risque d'inclure des variables particulièrement « étranges », sans justification théorique.

L'analyse de régression est un outil puissant car il permet de trouver des relations parmi un grand nombre de données. Les statistiques nous donnent des standards objectifs pour évaluer ces relations à condition

1) de définir une bonne équation de régression, définissant les bonnes variables à étudier sur les bonnes données = estimer l'équation, plus important que le calcul statistique sous-jacent.
2) L'analyse établit seulement un cas circonstanciel : une relation entre 2 variables nous met dans la bonne direction mais n'est pas suffisante pour affirmer (convaincre). Le résultat doit pouvoir être reproduit ou tout du moins être constant avec d'autres résultats.

-