-

Les concepts élémentaires des statistiques

I. Les variables

Les variables sont quelques choses que nous mesurons, commandons, ou manœuvrons dans la recherche. Elles diffèrent dans le rôle que nous donnons dans notre recherche et dans le type de mesures qui peuvent être appliquées à elles.

Les variables indépendantes sont ceux qui sont manipulées tandis que les variables dépendantes sont seulement mesurées ou enregistrées.

Les notions de variable dépendante et variable indépendante s'appliquent la plupart du temps à la recherche expérimentale où quelques variables sont manipulées, celles- ci dépendent de la manipulation ou des conditions expérimentales, c'est-à-dire, qu'elles dépendent de ce que le sujet fera dans la réponse. D'autres variables sont "indépendantes" des modèles initiaux de réaction, des dispositions, des intentions, etc… des sujets.

Exemple : si dans une expérience, des mâles sont comparés aux femelles concernant leur taux de globules blanches, le genre pourrait s'appeler la variable indépendante et les globules blanches, la variable dépendante.

II. Les échelles de mesure

Les variables diffèrent par deux aspects de mesures:

- "à quel point" elles peuvent être mesurées
- combien d'information mesurable leur échelle de mesure peut fournir.

Il existe une certaine erreur de mesure (impliquée dans chaque mesure) qui détermine la "quantité de l'information" qu'on peut obtenir. Le facteur qui détermine la quantité de l'information qui peut être fournie par une variable est son "type d'échelle de mesure." Les

variables sont classifiées selon leurs échelles de mesures en quatre types : variable nominale, nombre ordinal, intervalle ou de rapport.

a. **Les variables nominales** : elles répondent à une classification qualitative. Elles peuvent être mesurées seulement en termes d'appartenance à quelques catégories distinctives, que nous ne pouvons pas mesurer.

   Par exemple: Deux individus sont différents en termes de variable (par exemple, ils sont de race différente), mais nous ne pouvons pas dire qu'un individu a plus de qualité représentée par la variable. Les exemples typiques des variables nominales sont le genre, la race, la couleur, lieu d'origine, etc...

b. **Les variables ordinales** : nous pouvons ordonner ce que nous mesurons en termes de la qualité représentée par la variable

   Un exemple typique: d'une variable ordinale est le statut socio-économique des familles. Par exemple, nous savons que le niveau moyen-supérieur est plus haut que le niveau moyen, mais nous ne pouvons pas dire qu'il est, par exemple, de 18% plus haut. **Exemple :** nous pouvons dire que la mesure nominale fournit moins d'informations que la mesure ordinale, mais nous ne pouvons pas dire "combien moins" ou comment cette différence compare le nombre ordinal et les échelles d'intervalle.

c. **Les variables d'intervalle** : nous pouvons, non seulement, ordonner les items qui sont mesurés, mais également mesurer et comparer les tailles des différences entre elles.

   Exemple : la température mesurée en degrés Celsius constitue une échelle d'intervalle. Nous pouvons dire qu'une température de 40 degrés est plus haute qu'une température de 30 degrés, et qu'une augmentation de 20 à 40 degrés est deux fois plus qu'une augmentation de 30 à 40 degrés.

d. **Les variables de rapport :** elles sont très semblables aux variables d'intervalle ; en plus de toutes les propriétés des variables d'intervalle, elles comportent un point nul absolu identifiable, ainsi elles tiennent compte des rapports tels que x est deux fois davantage que les exemples typiques de y.

   Les échelles d'intervalle n'ont pas la propriété des échelles de rapport. La plupart des procédures statistiques d'analyse de données ne distinguent pas les propriétés de l'intervalle et du rapport des échelles de mesure.

III. Les relations entre les variables

Indépendamment de leur type, deux variables ou plus peuvent être reliées si dans un échantillon d'observations, les valeurs de ces variables sont distribuées d'une façon cohérente. En d'autres termes, les variables sont reliées, si leurs valeurs correspondent systématiquement entre eux pour ces observations.

Exemples :

- La taille est liée au poids parce que les individus grands sont généralement plus lourds que les petits individus.
- Le Q.I. est lié au nombre d'erreurs dans un essai, les gens avec un plus haut Q.I font peu d'erreurs.

L'importance de la relation entre les variables

D'une manière générale, le but final de chaque recherche ou analyse scientifique est de trouver des relations entre les variables. L'avancement de la science implique de trouver de nouvelles relations entre les variables. C'est la recherche de corrélation qui mesure de telles relations avec la façon la plus franche.

Les statistiques ne font que nous aider à évaluer les relations entre les variables. En fait, toutes les centaines de procédures des statistiques peuvent être interprétées en termes d'évaluation de divers genres de relations inter-variables.

Les deux propriétés pour chaque relation entre les variables

Les deux propriétés formelles les plus élémentaires de chaque relation entre les variables sont la grandeur de la relation (ou "taille") et son sérieux (ou "exactitude").

a. ***Grandeur*** **"taille".** Il est beaucoup plus facile de comprendre et mesurer la grandeur que la fiabilité.

   Par exemple : si chaque mâle dans un échantillon s'avère etre plus grand de taille que n'importe quelle femelle dans l'échantillon, nous pourrions dire que l'importance de la relation entre les deux variables (genre et taille) est très haute dans cet échantillon.

b. ***Fiabilité*** **"exactitude".** La fiabilité d'une relation est un concept beaucoup moins intuitif, mais toujours extrêmement important. Elle concerne la représentativité du résultat trouvé dans un échantillon spécifique par rapport à la population entière.

En d'autres termes, elle indique qu'il est probable, c'est qu'une relation semblable serait trouvée si l'expérience serait répétée avec d'autres échantillons tirés de la même population.

La fiabilité d'une relation entre les variables observées dans un échantillon peut être quantitativement estimée et représentée en utilisant une mesure standard (techniquement appelée niveau de p (probabilités) ou le niveau de signification statistique).

La force et la fiabilité d'une relation entre les variables

La force et la fiabilité sont deux dispositifs différents des rapports entre les variables. Cependant, elles ne sont pas totalement indépendantes. En général, dans un échantillon d'une taille particulière, plus l'importance de la relation entre les variables est grande, plus est fiable la relation.

Pourquoi des relations plus fortes entre les variables sont-elles plus significatives ?

L'importance et la signification d'une relation semblent être étroitement liées ; nous pourrons calculer la signification de la grandeur et vice-versa. Cependant, cela est vrai seulement, si la dimension de l'échantillon est gardée constante, parce que la relation pourrait être fortement significative ou non significative du tout, selon la dimension de l'échantillon.

Pourquoi la signification d'une relation entre les variables dépend de la taille de l'échantillon ?

La signification dépend, la plupart du temps, de la dimension de l'échantillon. Dans les échantillons très grands, même des relations très petites entre les variables sont significatives ; tandis que dans des échantillons très petits, même des relations très grandes ne peuvent pas être considérées fiables (significative).

S'il y a très peu d'observations, il y aurait également respectivement peu de combinaisons possibles des valeurs des variables, et la probabilité d'obtenir par hasard une combinaison de ces valeurs indicatives d'une relation forte est relativement haute.

Exemple: "bébé garçons par rapport au bébé filles ". Il y a deux hôpitaux ; dans le premier, 120 bébés naissent chaque jour, dans l'autre, seulement 12. En moyenne, le rapport des bébés

garçons et des bébés filles nées chaque jour dans chaque hôpital est de 50/50. Cependant, pendant un jour, dans un de ces hôpitaux il y a deux fois autant de bébés filles que des bébés garçons. Dans quel hôpital ça ce produit ?

La réponse est évidente pour un statisticien, ça se produit dans le petit hôpital. La raison de ceci, est que la probabilité d'une déviation aléatoire d'une telle taille diminue avec l'augmentation de la dimension de l'échantillon.

Pourquoi les petites relations peuvent-elles être significatives seulement dans de grands échantillons ?

Si un rapport entre les variables est petit, alors il n'y a aucune manière d'identifier une telle relation dans une étude à moins que l'échantillon de recherche soit grand. Même si notre échantillon est "parfaitement représentatif" l'effet ne sera pas statistiquement significatif si l'échantillon est petit.

Si une relation est très grande, l'effet peut s'avérer fortement significatif même dans une étude basée sur un échantillon très petit. Plus la relation entre les variables est petite, plus la dimension de l'échantillon devrait être grande pour prouver la signification.

Exemple : imaginer combien de jets en l'air seraient nécessaires pour montrer qu'une pièce de monnaie est asymétrique, si sa polarisation était seulement de .000001%.

IV. Le principe général commun de la plupart des tests statistiques

Les tests statistiques représentent un rapport de variation expliquée à la variation totale. Une fois nous avons calculé la mesure d'une relation entre deux variables ; la question serait : de combien cette relation est-elle significative?

Afin de déterminer le niveau d'importance statistique, nous avons besoin d'une fonction qui représente le rapport entre la "grandeur" et la "signification" des relations entre deux variables, selon la dimension de l'échantillon. La fonction dont nous avons besoin, nous indiquerait la relation d'une grandeur donnée d'un échantillon d'une taille donnée. Cette fonction nous donnerait le niveau de la signification (p), et elle nous indiquerait la probabilité de l'erreur impliquée. La représentation graphique de cette fonction de probabilité a dans la

plupart des cas, une forme particulière que nous pouvons employer pour déterminer les niveaux de signification de nos résultats dans les échantillons d'une taille particulière.

La plupart de ces fonctions sont liées à un type général de fonction qui s'appelle la fonction normale.

V. La distribution normale

La distribution de beaucoup de tests statistiques est normale ou suit une certaine forme qui peut être dérivée de la distribution normale. La distribution normale représente une des "vérités élémentaires empiriquement vérifiées au sujet de la nature générale de la réalité et son statut peut être comparé à celui des lois fondamentales des sciences naturelles.

La forme exacte de la distribution normale (la courbe en cloche) est définie par une fonction qui a seulement deux paramètres : la moyenne et l'écart type. Une propriété caractéristique de la distribution normale est que 68% de toutes ses observations font partie d'une marge de la moyenne ±1l'écart type, et que ±2 écarts type inclut 95% des points. En d'autres termes, dans une distribution normale, des observations qui ont une valeur normalisée de moins de -2 ou plus de +2 ont une fréquence relative de 5% ou moins. La valeur standardisée signifie qu'une valeur est exprimée en termes de sa différence à la moyenne, divisé par l'écart type.

Exemple : si on écrit la valeur de Z (valeur normalisée) de 4, la probabilité associée sera moins que .0001, car, dans la distribution normale, presque toutes les observations (plus de 99.99%) font partie de la marge des ±4 écart-type.

L'utilisation de la distribution normale dans l'induction statistique

Si la dimension de l'échantillon est assez grande, les résultats sont normalement distribués, par suite, nous pouvons calculer avec précision, la probabilité d'obtenir des résultats représentant de divers niveaux de déviation par rapport à la moyenne 0 de la population.

Si une probabilité calculée est basse, nous concluons que notre résultat diffère de l'hypothèse nulle ; l'hypothèse nulle a été considérée seulement pour "des raisons techniques" comme un repère contre lequel les résultats empiriques ont été évalués.

Tous les tests statistiques sont-ils normalement distribués ?

La plupart des tests sont basés directement sur la distribution normale ou sur les distributions dérivés de la normale, comme *t*, *F*, ou *Chi* carré. Typiquement, ces tests exigent que les variables analysées, elles-mêmes soient normalement distribuées dans la population, c.-à-d., elles obéissent à la "prétention de normalité."

Beaucoup de variables observées réellement sont normalement distribuées, ceci induit que la distribution normale représente "un dispositif général" de la réalité empirique. Le problème se produit quand on essaye d'employer un test de la distribution normale pour analyser les données des variables qui elles-mêmes ne sont pas normalement distribuées. Dans ces cas-ci nous avons deux choix :

1. nous pouvons employer un test "non paramétrique" alternatif (ou un test de libre distribution) ; mais c'est souvent incommode parce que de tels tests sont en général moins puissants et moins flexibles en terme de type de conclusions qu'ils peuvent fournir.
2. Dans beaucoup de cas, nous pouvons encore employer le test de la distribution normale si la taille de nos échantillons est assez grande. Cette dernière option est basée sur un principe extrêmement important qui est en grande partie responsable de la popularité des tests qui sont basés sur la fonction normale.

Principe de la loi centré réduite : A mesure que la dimension de l'échantillon augmente, la forme de la distribution de prélèvement (distribution d'une statistique de l'échantillon) se rapproche de la forme normale, même si la distribution de la variable en question n'est pas normale. Cependant, en augmentant les dimensions de l'échantillon, la forme de la distribution de prélèvement devient normale. Noter que pour n=30, la forme de cette distribution est "presque" normale.

Quelles sont les conséquences de la violation de la normalité ?

Les études de Monte Carlo ont analysé les tests basés sur la distribution normale pour déterminer à quel point elles sont sensibles aux violations de l'acceptation de la distribution normale. Les conclusions de ces études stipulent que les conséquences de telles violations sont moins graves qu'on ne l'a pensé précédemment. Ceci a permis d'augmenter la popularité globale des tests statistiques dans tous les domaines de la recherche.