1 Introduction à la Statistique non-paramétrique

Une grande partie du domaine des statistiques et des méthodes statistiques est consacrée aux données dont la distribution est connue. Les échantillons de données dont nous connaissons déjà ou pouvons facilement identifier la distribution sont appelés données paramétriques. Souvent, paramétrique est utilisé pour désigner des données tirées d'une distribution gaussienne d'usage courant. Les données dont la distribution est inconnue ou ne peuvent pas être facilement identifiées sont appelées non paramétriques.

Dans le cas où vous travaillez avec des données non paramétriques, des méthodes statistiques non paramétriques spécialisées peuvent être utilisées pour éliminer toutes les informations sur la distribution. En tant que telles, ces méthodes sont souvent appelées méthodes sans distribution.

Les méthodes statistiques ont les fonctions principales suivantes: (1) la conception d'hypothèses et de procédures expérimentales et la collecte de données; (2) la présentation synthétique des données pour une compréhension facile, claire et significative; et (3) l'analyse des données quantitatives pour fournir des conclusions valables sur les phénomènes observés. Pour ces trois fonctions principales, deux types de méthodes sont généralement appliqués: paramétrique et non paramétrique.

1.1 Différences entre la statistique paramétrique et la statistique non-parmétrique

1.1.1 La statistique paramétrique

1- La statistique paramétrique est le cadre "classique"de la statistique. Le modèle statistique y est décrit par un nombre ni de paramètres. Typiquement M= fP ; est le modèle statistique qui décrit la distribution des variables aléatoires observées.

2- Les méthodes paramétriques sont basées sur une distribution normale ou gaussienne, caractérisée par la moyenne et l'écart type. La distribution des résultats est symétrique autour de la moyenne, avec 95% des résultats à moins de deux écarts-types de la moyenne.

3- Les statistiques paramétriques sont utilisées avec des données d'intervalle continues qui montrent l'égalité des intervalles ou des différences.

1.1.2 La statistique non-parmétrique

1- En statistique non param etrique, le modèle n'est pas décrit par un nombre ni de paramètres. Divers cas de gures peuvent se présenter, comme par exemple : On s'autorise toutes les distributions possibles, i.e. on ne fait aucune hypothèse sur la forme/nature/type de la distribution des variables aléatoires. On travaille sur des espaces fonctionnels, de dimension infinie.

Exemple : les densités continues sur [0; 1], ou les densités monotones sur R. Le nombre de paramètres du modèle n'est pas x e et varie (augmente) avec

le nombre d'observations. Le support de la distribution est discret et varie (augmente) avec le nombre d'observations.

1- Les statistiques non paramétriques ne sont pas basées sur de telles distributions de probabilité paramétrées ni même sur des hypothèses concernant la distribution de probabilité des données.

2- Les méthodes non paramétriques sont appliquées aux données ordinales, telles que les données d'échelle de Likert impliquant la détermination de plus grand ou plus petit , c'est-á-dire le classement des données.

1.2 Données paramétriques et données non paramétriques

1.2.1 Données paramétriques

Les données paramétriques sont un échantillon de données tirées d'une distribution de données connue. Cela signifie que nous connaissons déjà la distribution ou que nous avons identifié la distribution, et que nous connaissons les paramètres de la distribution. Souvent, paramétrique est un raccourci pour des données à valeur réelle tirées d'une distribution gaussienne. Il s'agit d'un raccourci utile, mais strictement ce n'est pas tout à fait exact.

Si nous avons des données paramétriques, nous pouvons utiliser des méthodes paramétriques. Poursuivant avec le raccourci de la signification paramétrique gaussienne. Si nous avons des données paramétriques, nous pouvons exploiter toute la suite de méthodes statistiques développées pour les données supposant une distribution gaussienne, telles que:

Statistiques récapitulatives.

Corrélation entre les variables.

Tests de signification pour comparer les moyennes.

En général, nous préférons travailler avec des données paramétriques, et même aller jusqu'à utiliser des méthodes de préparation de données qui rendent les données paramétriques, telles que des transformations de données, afin de pouvoir exploiter ces méthodes statistiques bien comprises.

1.2.2 Données non paramétriques

Les données qui ne correspondent pas à une distribution connue ou bien comprise sont appelées données non paramétriques. Les données peuvent être non paramétriques pour de nombreuses raisons, telles que:

Les données n'ont pas de valeur réelle, mais plutôt des nombres ordinaux, des intervalles ou une autre forme.

Les données ont une valeur réelle mais ne correspondent pas à une forme bien comprise.

Les données sont presque paramétriques mais contiennent des valeurs aberrantes, plusieurs pics, un décalage ou une autre caractéristique.

Il existe une suite de méthodes que nous pouvons utiliser pour les données non paramétriques appelées méthodes statistiques non paramétriques. En fait, la plupart des méthodes paramétriques ont une version non paramétrique équivalente.

En général, les résultats des méthodes non paramétriques sont moins puissants que leurs homologues paramétriques, notamment parce qu'ils doivent être généralisés pour fonctionner pour tous les types de données. Nous pouvons toujours les utiliser pour l'inférence et faire des déclarations sur les découvertes et les résultats, mais elles n'auront pas le même poids que les affirmations similaires avec des méthodes paramétriques. Les informations sur la distribution sont supprimées.

1.3 Domaines d'utilisation de la statistique non paramétrique

Exemple d'observations mesur ees

On observe des donn ees quantitatives. Questions : Peut-on raisonnablement supposer que les observations suivent une loi normale ? (par exemple pour faire des tests sur la moyenne). Rep : Tests de normalité.

Combien de modes possède cette distribution ? Rep : Estimation de densité.

Exemples de contextes d'utilisation

- Quand on n'arrive pas a ajuster correctement les observations avec une distribution paramétrique.

- Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle.

- Quand on ne sait pas combien de composantes on veut mettre dans un mélange.

- Quand le nombre de variables est trop grand (problème de grande dimension) et qu'un modèle paramétrique est non utilisable car il aurait de toutes façons trop de paramètres,...etc.

Avantages

- Moins d'a priori sur les observations.

- Modèles plus généraux, donc plus robustes au modèle.

Inconvénients

- Vitesses de convergence plus lentes = il faut plus de données pour obtenir une précision equivalente.