Synthèse du cours de Statistiques

Statistiques univariées

Vocabulaire et classification des variables

- **Variable discrète** : prend un nombre dénombrable de valeurs (exemple : nombre d'élèves ayant une note donnée).
- **Variable continue** : peut prendre toutes les valeurs dans un intervalle (exemple : poids mesuré entre 60 kg et 70 kg).
- **Variables qualitatives et quantitatives** :
  - Qualitative : regroupe des modalités non numériques (exemple : couleurs).
  - Quantitative : associée à des valeurs numériques.

Représentation des données

- **Diagramme en bâtons** : pour les variables discrètes.
- **Histogramme** : pour les variables continues, où l'aire des rectangles correspond à l'effectif.
- **Fonction de répartition** : permet de répondre à des questions comme "combien d'élèves ont une note inférieure ou égale à 4 ?".

Caractéristiques de position

- **Mode** : valeur la plus fréquente

  (Exemple :

  Supposons qu'on ait les notes de 10 élèves : 4, 6, 7, 4, 8, 4, 5, 6, 7, 4,

  La note 4 est répétée 4 fois. Les autres notes (6, 7, 8, 5) sont moins fréquentes : 4 est le mode de cet ensemble de données).
- **Médiane** : valeur qui sépare la série en deux parties égales. Exemple : si 10 élèves passent un examen, la médiane est la moyenne des notes des 5e et 6e élèves.

- **Moyenne** : somme des valeurs divisée par leur nombre. Exemple :

$$\bar{x} = \frac{\sum(x_i \cdot n_i)}{N}.$$

Caractéristiques de dispersion

- **Étendue** : différence entre la valeur maximale et minimale.
- **Variance et écart-type** :

$$\mathrm{Var}(x) = \frac{\sum(x_i - \bar{x})^2 n_i}{N}.$$

  L'écart-type est la racine carrée de la variance (σ), indiquant l'écart moyen autour de la moyenne.

Statistiques bivariées

Analyse des relations entre deux variables

- **Distributions conditionnelles et marginales** :
  - **Marginale** : répartition de chaque variable indépendamment (somme sur les lignes ou colonnes).
  - **Conditionnelle** : répartition d'une variable en fixant une modalité de l'autre.

Mesures de dépendance

Objectif : vérifier si deux variables sont indépendantes ou non. La statistique du χ^2 est donnée par :

$$D_{\chi^2} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{j\cdot} f_{\cdot k})^2}{f_{j\cdot} f_{\cdot k}}.$$

Les coefficients suivants mesurent la dépendance entre deux variables :

- $\phi = \sqrt{\frac{D_{\chi^2}}{n}}$: intensité de la relation.
- V de Cramér : plus précis pour les tableaux non carrés.

L'interprétation de V :

- $V < 0.1$: relation faible ou nulle.
- $V > 0.3$: relation forte.

Régression linéaire

Nuage de points

Chaque individu est représenté par un point (x_i, y_i). Le **point moyen** est donné par $(\bar{x}, \bar{y})$, centre de gravité du nuage.

Méthode des moindres carrés

L'objectif est de minimiser la somme des carrés des distances verticales entre les points et la droite de régression :

$$a = \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}, \quad b = \bar{y} - a\bar{x}.$$

La covariance est calculée par :

$$\mathrm{Cov}(x, y) = m(xy) - \bar{x}\bar{y}.$$

Coefficient de corrélation linéaire (r)

La mesure de la relation linéaire entre x et y est donnée par :

$$r = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}.$$

$r = 1$ ou $r = -1$ indique une relation parfaitement linéaire.

Coefficient de détermination (R^2)

Il mesure la proportion de la variance de y expliquée par x. Ce coefficient équivaut au coefficient de corrélation linéaire au carré (r^2) :

$$R^2 = \frac{\mathrm{Var}(ax + b)}{\mathrm{Var}(y)}.$$

$$R^2 = (r^2)$$

Tests statistiques

Hypothèses et p-value

L'**hypothèse nulle** (H_0) est la proposition initiale à tester (ex. : indépendance des variables). La **p-value** est la probabilité d'erreur si H_0 est rejetée. On rejette H_0 si $p \leq \alpha$ (généralement $\alpha = 0.05$).

Tests sur régression

- **Test t de Student** : vérifie l'effet individuel des variables explicatives.
- **Test F de Fisher (ANOVA)** : analyse globale de l'effet des variables explicatives.

Tests de normalité

Les tests de normalité, comme le test de Shapiro-Wilk, permettent de vérifier si une série suit une loi normale.Si $p > 0.05$, on ne rejette pas H_0 et la série peut être supposée normale.

Test du Chi-2

Le test du Chi-2 permet de tester l'indépendance entre deux variables qualitatives à partir d'un tableau de contingence. Voici les étapes principales :

1. **Hypothèse nulle (H_0)** : Les deux variables sont indépendantes.
2. **Statistique du test** : La distance du Chi-2 (D_{χ^2}) est calculée par la formule :
$$D_{\chi^2} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{j\bullet} f_{\bullet k})^2}{f_{j\bullet} f_{\bullet k}}$$
où f_{jk} est la fréquence observée, et $f_{j\bullet}$ et $f_{\bullet k}$ sont les marges des fréquences.
3. **Quantile et décision** :
   - On détermine $d = (J-1)(K-1)$, le degré de liberté.
   - Au seuil α (souvent 5%), on compare D_{χ^2} au quantile $q_{1-\alpha}$ d'une loi χ^2_d.
   - Si $D_{\chi^2} \geq q_{1-\alpha}$, on rejette H_0, concluant que les variables sont dépendantes.

Exemple : Avec $D_{\chi^2} = 11.6$ pour $d = 1$, et $q_{0.95} = 3.84$, on conclut à une dépendance car $11.6 > 3.84$.

ANOVA à un Facteur

L'ANOVA (Analysis of Variance) à un facteur évalue si une variable qualitative (facteur) explique une différence significative entre les moyennes d'une variable quantitative.

Utilisation :

- En L1 on ne passe que par la p-value. La **p-value** est utilisée pour déterminer si les moyennes des différents groupes ou niveaux du facteur étudié sont statistiquement différentes les unes des autres. La p-value est calculée à partir du rapport entre la variance expliquée par les groupes (variance inter-groupes) et la variance résiduelle (variance intra-groupe).

  - Une p-value inférieure à un seuil prédéfini (souvent $\alpha = 0,05$) suggère que les différences observées entre les groupes ne sont pas dues au hasard, conduisant au rejet de H_0. Cela indique qu'au moins un des groupes diffère significativement des autres.
  - Une p-value supérieure au seuil indique qu'il n'y a pas suffisamment de preuves pour rejeter H_0, et les moyennes des groupes sont considérées comme statistiquement similaires.

2

Ces outils sont essentiels pour tester des hypothèses en statistique. Le test du Chi-2 explore les relations entre variables qualitatives, tandis que l'ANOVA examine les effets d'un facteur qualitatif sur une variable quantitative.Et l'ANOVA sur une régréssion linéaire montre l'effet d'une quantitative sur une quantitative
.