1 Statistiques Descriptives

1.1 Variable Discrète

Définition 1 Soit x une variable discrète, à valeurs dans $V_x = \{\xi_1, \ldots, \xi_K\}$ avec $\xi_1 < \ldots < \xi_K$. Une distribution de n observations de x peut être représentée sous forme d'un tableau de fréquences où figure pour chaque modalité de ξ_K de x, le nombre n_k d'observations ayant la valeur ξ_K.La fréquence relative est $f_k = \frac{n_k}{n}$. La fréquence cumulée est $F_k = \sum_{j=1}^{k} f_j$

1.2 Variable Continue

Définition 2 Si x est continue, on partitionne le domaine de définition de x en K classes. On prendra la règle de Sturges pour calculer K. $K = 1 + \frac{10}{3}\log_{10} n$.

Pour une variable discrète, on va utiliser un diagramme en baton, alors que pour une variable continue, on va utiliser un histogramme, où l'aire de chaque rectangle est proportionnel à l'effectif.

1.3 Fonction de répartition empirique

Définition 3 La fonction de répartition empirique est :

$$\hat{F} : \mathbb{R} \to [0,1]$$

$$x \to \frac{1}{n}\sharp\{i : x_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} 1_{[x_i,+\infty[}(x)$$

1.4 Indicateur de tendance centrale

Définition 4 Moyenne empirique : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Dans le cas des classes, on a : $\overline{x} = \frac{1}{n}\sum_{k=1}^{K} n_k c_k$

Proposition 1 La somme des écarts à la moyenne empirique est nulle : $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x}$ La moyenne empirique est la valeur la plus proche au sens de la somme des carrés des écarts.

Définition 5 Fractile d'ordre α

$$f_\alpha = (1-\gamma)x_{(j)} + \gamma x_{(j+1)}$$

avec $j = \lfloor \alpha(n+1) \rfloor$ et $\gamma = \alpha(n+1) - j$

1.5 Indicateur de dispertion

Définition 6 La variance empirique : $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$

La variance empirique corrigée : $s^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}s^2$

1.6 Boite à Moustaches

Définition 7 Il s'agit d'un graphique formé d'une boite délimitée par le premier et le troisième quartile, où figure aussi la médiane ($f_{0.5}$) Des segments de droites s'étendent de part et d'autre de la boite jusqu'au point le plus extrême à une distance inférieure à 1.5H. Les autres points sont représentés individuellement. H est la hauteur de la boite.

2 Rappel de Probabilité

2.1 Probabilité

Proposition 2 Propriétés des probabilitées :
- $P(\emptyset) = 0$;
- $P(\overline{A}) = 1 - P(A)$;
- $A \subset B \Rightarrow P(A) \leq P(B)$;
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
- $P(\cup A_i) \leq \sum P(A_i)$;

Proposition 3 A et B sont indépendant si $P(A \cap B) = P(A)P(B)$. n événements $A_1, \ldots, A_n$ sont indépendant si pour toute partie I de $[1,n]$: $P(\cap_{i\in I} A_i) = \Pi_{i \in I} P(A_i)$.

2.2 Variables Aléatoires

Proposition 4 Propriétés de la densité
- $\forall x : f(x) \geq 0$
- $\int_{\mathbb{R}} f(x)dx = 1$
- $P(x < X < x + \Delta x) = f(x)\Delta x + o(\Delta x)$

2.3 Fonction de répartition

Définition 8 $F : \mathbb{R} \to [0,1]$ definie par $\forall x \in \mathbb{R} : F(x) = P(X \leq x)$.

Proposition 5 Propriété de la fonction de répartition
- F est croissante et continue à droite.
- $\lim_{x\to-\infty} F(x) = 0$
- $\lim_{x\to+\infty} F(x) = 1$
- $P(a < X \leq b) = F(b) - F(a)$
- Si X est continue, alors F est dérivable et on a $F(x) = \int_{-\infty}^{x} f(t)dt$

2.4 Espérance mathématique

Proposition 6 Propriétés de l'espérance
- Si $X \geq 0$, alors $E(X) \geq 0$
- $E(a) = a$
- $E(aX + bY) = aE(x) + bE(Y)$

2.5 Variance

Proposition 7 Propriétés de la variance
- $E((X-a)^2) = Var(X) + (E(X) - a)^2$
- $Var(X) = E(X^2) - (E(X))^2$
- $Var(aX + b) = a^2 Var(X)$
- $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$

2.6 Autres moments

Définition 9
- Moment centré d'ordre k : $\mu_k = E([X - E(X)]^2)$
- Moment non centré d'ordre k : $m_k = E(X^k)$

2.7 Convergence Stochastique

Définition 10 La suite (X_n) converge en probabilité vers a si $\forall \varepsilon > 0$ et $\eta > 0$, $\exists n_0$ tel que $\forall n > n_0$

$$P(|X_n - a|) > \varepsilon) < \eta$$

.On note $(X_n) \xrightarrow{P} a$. La convergence en probabilité de X_n vers X revient à étudier la convergence de $X - X_n$ vers 0. Condition : Si $E(X_n) \to a$ et $Var(X_n) \to 0$, alors $X_n \xrightarrow{P} a$.

5.4.2 Propriétés

Les estimateurs $\hat{\theta}_m$ obtenus par la méthode des moments sont :
- convergents : $\hat{\theta}_m \xrightarrow{P} \theta$
- asymptotiquement gaussiens :

$$\sqrt{n}(\hat{\theta}_m - \theta) \xrightarrow{L} \mathcal{N}(0, \sigma_m^2)$$

6 Intervalles de confiance

6.1 Définition

$X_1, \ldots, X_n$ étant un échantillon i.i.d dont la loi parente X a une distribution dépendante du paramètre réel θ, α un réel quelconque fixé appartenant à $]0,1[$, on appelle intervalle de confiance pour le paramètre θ au niveau de confiance $1-\alpha$ tout intervalle $[T_1, T_2]$ où T_1 et T_2 sont deux statistiques vérifiant $P(T_1 \leq \theta \leq T_2) = 1 - \alpha$.

6.2 Construction

- Choisir un estimateur T de θ dont on connait la loi de probabilité en fonction de θ ;
- Déterminer $f(T,\theta)$ dont la loi de probabilité ne dépend plus de θ : la fonction *pivotale* ;
- Si c'est possible, en déduire $P(T_1 \leq \theta \leq T_2) = 1 - \alpha$ où T_1 et T_2 sont des statistiques fonctions de T.

6.3 Variable normale

Hypothèse : $X_1, \ldots, X_n$, $X_i \sim \mathcal{N}(\mu, \sigma^2)$, niveau de confiance $1-\alpha$, intervalle bilatéral.

6.3.1 Moyenne

1. Si σ est connue, on a $T = \overline{X}$ (estimateur) et $\theta = \mu$, on sait que $f(T,\theta) = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$.

   Comme $P(u_{\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}) = 1-\alpha$, on obtient comme I.C $[\overline{X} - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}]$. En unilatéral, on écrit $P(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\alpha}) = 1-\alpha$ et on obtiendrait $[\overline{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha}, +\infty]$.

2. Sinon, on remplace σ par S^* pour remarquer que

$$\frac{\overline{X}-\mu}{\frac{S^*}{\sqrt{n}}} = \frac{\left.\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right\} \sim \mathcal{N}(0,1)}{\left.\sqrt{\frac{(n-1)\frac{S^{*2}}{\sigma^2}}{n-1}}\right\} \sim \sqrt{\frac{\chi^2_{n-1}}{n-1}}} \sim \tau_{n-1} \text{ (pleure pas).}$$

   On obtient alors $P(-t_{n-1;1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{S^*}{\sqrt{n}}} \leq t_{n-1;1-\frac{\alpha}{2}}) = 1-\alpha$ d'où I.C=$[\overline{X} - \frac{S^*}{\sqrt{n}}t_{n-1;1-\frac{\alpha}{2}}, \overline{X} + \frac{S^*}{\sqrt{n}}t_{n-1;1-\frac{\alpha}{2}}]$.

6.3.2 Variance

1. Si μ est connue, l'estimateur T est l'estimateur de maximum de vraisemblance $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$, d'où $f(T,\theta) = \frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{n}(\underbrace{\frac{X_i - \mu}{\sigma}}_{\sim\mathcal{N}(0,1)})^2 \sim \chi^2_n$ et on utilise la relation $P(\frac{n\hat{\sigma}^2}{\chi^2_{n;1-\frac{\alpha}{2}}} < \sigma^2 < \frac{n\hat{\sigma}^2}{\chi^2_{n;\frac{\alpha}{2}}})$.
2. Sinon, l'estimateur est S^{*2} et $f(T,\theta) = (n-1)\frac{S^{*2}}{\sigma^2} \sim \chi^2_{n-1}$.

   En utilisant

$$P(\chi^2_{n-1;\frac{\alpha}{2}} \leq (n-1)\frac{S^{*2}}{\sigma^2} \leq \chi^2_{n-1;1-\frac{\alpha}{2}}) = 1-\alpha$$

   on obtient $[(n-1)\frac{S^{*2}}{\chi^2_{n-1;1-\frac{\alpha}{2}}}, (n-1)\frac{S^{*2}}{\chi^2_{n-1;\frac{\alpha}{2}}}]$

6.4 Utilisation du maximum de vraisemblance

Soit θ un paramètre, $\hat{\theta}$ son estimateur de maximum de vraisemblance. Alors $\frac{\hat{\theta}-\theta}{\sqrt{\frac{1}{I_n}}} \xrightarrow{L} \mathcal{N}(0,1)$, ce qui constitue une fonction asymptotiquement pivotale pour θ.

7 Tests

7.1 Definitions

7.1.1 Hypothèses

Soit $X \sim P_\theta, \theta \in \Theta$. Une hypothèse est une partie H de $\theta : H \subset \Theta$. Elle est simple si H est un singleton, composite sinon. Un problème de test consiste en un choix entre deux hypothèses H_0 et H_1. Elles sont supposées complémentaires : $H_0 \cap H_1 = \emptyset$ et $H_0 \cup H_1 = \Theta$.

7.1.2 Règle de décision

Une règle de décision est une application $\varphi : \mathbb{R}^n \mapsto D$ qui à toute réalisation de l'échantillon $(x_1, \ldots, x_n)$ associe une décision $\varphi(x_1, \ldots, x_n)$ (choix de H_0 ou de H_1). Pour résoudre le test, on partitionne l'espace en deux sous-ensembles :
- $W = \varphi^{-1}(d_1)$ des réalisations pour lesquelles on rejette H_0 : c'est la région critique.
- $\overline{W} = \varphi^{-1}(d_0)$ des réalisations pour lesquelles on rejette H_1 : la région d'acceptation.

Remarque : La détermination des hypothèses et de la région critique doit se faire avant de connaitre le résultat de l'expérience.

7.1.3 Risques et puissance

- Le *risque de première espèce* est la probabilité de rejeter à tort l'hypothèse nulle. Soit $\forall\theta_0 \in H_0, \alpha_{\theta_0}(\varphi) = P_{\theta=\theta_0}((X_1, \ldots, X_n) \in W)$. Si H_0 est une hypothèse simple, on notera $\alpha(\varphi)$ ou α s'il n'y a pas d'ambiguite sur le test.
- La borne supérieure du risque de première espèce est le *niveau de signification* du test : $\alpha^*(\varphi) = \sup_{\theta_0 \in H_0} \alpha_{\theta_0}(\varphi)$
- Le *risque de seconde espèce* est la probabilité d'accepter à tort l'hypothèse nulle. $\forall\theta_1 \in H_1, \beta_{\theta_1}(\varphi) = P_{\theta=\theta_1}((X_1, \ldots, X_n) \in \overline{W})$.
- La quantité $1-\beta_{\theta_1}(\varphi)$ est appelée *puissance du test* (pour $\theta = \theta_1$) : c'est la probabilité d'accepter à juste titre H_1. La fonction $\theta_1 \mapsto 1-\beta_{\theta_1}(\varphi)$ est la fonction de puissance.

7.2 Méthode de Neyman–Pearson

On contrôle le risque de première espèce en le rendant inférieur à une valeur fixée à priori à un α^* donné. On restreint donc à la classe $\mathcal{C}(\alpha^*) = \{\varphi : \mathbb{R}^n \mapsto \{d_0, d_1\} | \alpha_{\theta_0}(\varphi) \leq \alpha^*, \forall\theta_0 \in H_0\}$. On cherche alors $\varphi^* \in \mathcal{C}(\alpha^*)$ telle que $1 - \beta_{\theta_1}(\varphi^*) = \max_{\varphi\in\mathcal{C}(\alpha^*)}(1-\beta_{\theta_1}(\varphi))\forall\theta_1 \in H_1$. On cherche donc le test le plus puissant pour un niveau de signification donné.

7.3 Cas particuliers

7.3.1 Hyp. simple *vs* Hyp. simple

On veut tester

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta = \theta_1 (\neq \theta_0) \end{cases}$$

Théorème 6 (Neyman et Pearson) La région critique optimale vérifie une relation de la forme $\frac{L(\theta_1; x_1, \ldots, x_n)}{L(\theta_0; x_1, \ldots, x_n)} > \lambda$, où λ se calcule par $P_{H_0}((X_1, \ldots, X_n) \in W) = \alpha^*$

7.3.2 Hyp. simple *vs* Hyp. composite

1. On veut résoudre

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \in E(\theta_0 \notin E) \end{cases} \Rightarrow \begin{cases} H_0 : & \theta = \theta_0 \\ H_i : & \theta = \theta_i (\theta_i \in E) \end{cases}$$

   Si la région critique W_i est indépendante de i et toujours égale à W, elle est optimale pour le test $\theta = \theta_0$ contre $\theta \in E$. L'erreur de seconde espèce sera une fonction de θ, mais sera toujours minimale pour α^* donné. Le test est donc dit *UPP* (Uniformément Plus Puissant).
2. Cas non *UPP* : si W_i n'est pas constant, la stratégie de Neyman–Pearson ne peut s'appliquer. On conserve la contrainte $\alpha = P_{H_0}((X_1, \ldots, X_n) <\in W) \leq \alpha^*$ fixé et on recherche une région critique en utilisant une fonction pivotale.

7.3.3 Hyp. composite *vs* Hyp. composite

1. Test *UPP*

   Définition 29 Une famille de lois L_θ est une famille à rapport de vraisemblance monotone s'il existe une statistique T telle que $\frac{L(\theta'; x_1, \ldots, x_n)}{L(\theta; x_1, \ldots, x_n)}$ soit une fonction croissante de t pour $\theta' > \theta$.

   Théorème 7 (Lehman) Si la famille des lois L_θ est à rapport de vraisemblance monotone, alors le problème de test

$$\begin{cases} H_0 : & \theta \leq \theta_0 \\ H_i : & \theta > \theta_0 \end{cases}$$

   admet un test *UPP* défini par une région critique de la forme $T > A$ où A est défini par $P_{\theta=\theta_0}(T > A) = \alpha^*$.

2. Test non *UPP*

   Forme générale : X suit une loi dépendant de plusieurs paramètres dont le paramètre réel θ sur lequel porte le test :

$$\begin{cases} H_0 : & \theta = \theta_0 & \text{autres paramètres inconnus} \\ H_i : & \theta >, <, \neq \theta_0 & \text{autres paramètres inconnus} \end{cases}$$

   Stratégie :
   - Trouver une fonction pivotale de θ
   - Proposer une forme de région critique raisonnable
   - Déterminer cette région critique en prenant $\alpha = P_{H_0}(X \in W) = \alpha^*$.

8 Tests d'hypothèse : comparaison de 2 échantillons

8.1 Comparaison de deux échantillons gaussiens

Définition 30 (loi de Fisher) Si X et Y sont 2 variables aléatoires indépendantes de loi de proba respectives $\chi^2_{\nu_1}$ et $\chi^2_{\nu_2}$, alors la variable $\frac{X/\nu_1}{Y/\nu_2}$ suit une loi de Fisher à ν_1 et ν_2 degrés de liberté.

$$F_{\nu_1,\nu_2,\alpha} = \frac{1}{F_{\nu_2,\nu_1,1-\alpha}} \qquad \text{(voir les tables)}$$

8.1.1 Variances : test de Fisher

$$\begin{cases} H_0 : \sigma_X^2 = \sigma_Y^2 & (HC) \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 & (HC) \end{cases} \text{ ou } \begin{cases} H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1 & (HC) \\ H_1 : \frac{\sigma_X^2}{\sigma_Y^2} \neq 1 & (HC) \end{cases}$$

Régions critiques : $S_X^{*2}/S_Y^{*2} < F_{n-1,m-1,\alpha}$ ou $S_X^{*2}/S_Y^{*2} > F_{n-1,m-1,\alpha}$.

Test unilatéraux : $\begin{cases} H_0 : \sigma_X^2 = \sigma_Y^2 & (HC) \\ H_1 : \sigma_X^2 < \text{ (resp >) } \sigma_Y^2 & (HC) \end{cases}$

Régions critiques : $S_X^{*2}/S_Y^{*2} < F_{n-1,m-1,\alpha}$ (resp $S_X^{*2}/S_Y^{*2} > F_{n-1,m-1,\alpha}$).

8.1.2 Moyennes : test de Student

Estimateur de la variance commune à 2 populations gaussiennes $(N = m + n)$: $S^{*2} = \frac{1}{N-2}((n-1)S_X^{*2} + (m-1)S_Y^{*2})$.

Remarque : $\frac{(N-2)S^{*2}}{\sigma^2} = \frac{(n-1)S_X^{*2}}{\sigma^2} + \frac{(m-1)S_Y^{*2}}{\sigma^2} \sim \chi^2_{N-2}$.

$X \sim \mathcal{N}(\mu_X, \sigma^2)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$:

$$\begin{cases} H_0 : \mu_X = \mu_Y & (HC) \\ H_1 : \mu_X \neq \mu_Y & (HC) \end{cases}$$

Région critique :
- $\frac{\overline{X}-\overline{Y}}{S^*\sqrt{1/n+1/m}} < t_{N-2,\alpha/2}$
- $\frac{\overline{X}-\overline{Y}}{S^*\sqrt{1/n+1/m}} > t_{N-2,1-\alpha/2}$

8.2 Tests non paramétriques

8.2.1 Test de Wilcoxon

$$\begin{cases} H_0 : F_X = F_Y \\ H_1 : F_X > F_Y \end{cases}$$

Suite W_X : fusion des réalisations des deux échantillons et classement par ordre croissant.

Si les deux échantillons on la même distribution $(F_X = F_Y)$ alors $E(W_X) = \frac{n(n+m+1)}{2}$ et $Var(W_X) = \frac{nm(n+m+1)}{12}$.

Région critique : $W_X < \frac{n(n+m+1)}{2} - u_{1-\alpha^*}\sqrt{\frac{nm(n+m+1)}{12}}$

8.3 Échantillons appariés

Pour l'étude de couples, en règle général, on étudie la différence : $D_i = Y_i - X_i$.

$$\begin{cases} H_0 : m_e = 0 \\ H_1 : m_e > \text{ ou } < \text{ ou } \neq 0 \end{cases}$$