Probabilités & Statistique

Laurent GARDES

Université de Strasbourg

Année universitaire 2023 / 2024

Partie II : Statistique

Population versus échantillon

Définition

Une population est l'ensemble de tous les individus partageant certaines caractéristiques communes.

- Exemple : population mondiale dont les individus partagent plusieurs caractéristiques communes (taille (en cm), poids (en kg), etc.).
- Il est souvent impossible de mesurer les caractéristiques sur l'ensemble des individus de la population.
- On se restreint donc dans ce cas à un sous-ensemble de la population.

Population versus échantillon

Définition

Un échantillon de taille $n \in \mathbb{N} \setminus \{0\}$ est un sous-ensemble de n individus de la population.

- Le choix de ce sous-ensemble peut être fait de manière aléatoire ou en essayant de respecter les répartitions de certaines caractéristiques de la population (par exemple avoir le même pourcentage d'hommes et de femmes, de retraités et d'actifs, etc.).
- L'objectif principal de la statistique est l'étude de la distribution d'une caractéristique d'une population à partir d'un échantillon.

Modèle statistique

- On mesure la caractéristique d'intérêt (que l'on supposera réelle) sur les n individus de l'échantillon.
- On dispose alors des valeurs observées $\mathbf{x} = (x_1, \cdots, x_n)$.
- On propose un modèle statistique s'ajustant au mieux à l'échantillon observé.

Définition

Un modèle statistique est un vecteur aléatoire $\mathbf{X} = (X_1, \cdots, X_n) : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *où* $(\Omega, \mathcal{A}, \mathbb{P})$ *est un espace probabilisé.*

- L'ensemble Ω est en fait la population et $(x_1, \cdots, x_n)$ est une réalisation de $\mathbf{X}$.
- Sous cette forme, le modèle est encore trop général pour permettre l'étude de la caractéristique.

Modèle statistique

• On supposera souvent que les variables aléatoires $X_1, \cdots, X_n$ sont indépendantes et de même loi $\mathbb{P}_X$ qu'une variable aléatoire $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

• La loi $\mathbb{P}_X$ est donc censée être la loi suivie par la caractéristique de la population. Cette loi est évidemment inconnue.

• On pourra supposer qu'elle appartient à une famille de loi paramétrique $\{\mathbb{P}_\theta;\ \theta \in \Theta\}$ (i.e. $\mathbb{P}_X = \mathbb{P}_{\theta_0}$). L'objectif sera alors l'estimation du paramètre θ_0.

• Dans toute la suite, on se placera dans le cadre d'un échantillon issu de variables aléatoires indépendantes et de même loi $\mathbb{P}_X \in \{\mathbb{P}_\theta;\ \theta \in \Theta\}$.

Définition

Un estimateur est une statistique observable c'est-à-dire une variable aléatoire $T(\mathbf{X})$ où $T : \mathbb{R}^n \to \mathbb{R}$ est une fonction mesurable connue. La valeur observée de l'estimateur est le réel $T(\mathbf{x})$ que l'on appelle estimation.

Exemple – Si $\mathbb{P}_X$ est une loi exponentielle de paramètre $\lambda > 0$, un estimateur de $\theta_0 = 1/\lambda$ est

$$\widehat{\theta}_n := T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

L'estimation de θ_0 est donc la valeur $(x_1 + \cdots + x_n)/n$.

Définition

Un estimateur $\widehat{\theta}_n$ de θ_0 est dit sans biais si $\mathbb{E}[\widehat{\theta}_n] = \theta_0$. Il est dit sans biais de variance minimale s'il est sans biais et si

$$\mathrm{Var}(\widehat{\theta}_n) = \min\{\mathrm{Var}(\check{\theta}_n);\ \check{\theta}_n \in \mathrm{SB}(\theta_0)\},$$

où $\mathrm{SB}(\theta_0)$ est l'ensemble des estimateurs sans biais de θ_0.

- Il existe plusieurs méthodes permettant d'obtenir un estimateur.

Nous allons en détailler trois :

1. La méthode des moments;
2. La méthode du maximum de vraisemblance;
3. L'estimation par intervalle de confiance.

- On suppose que $\Theta \subset \mathbb{R}^p$ avec $p \in \mathbb{N} \setminus \{0\}$.
- On suppose que pour tout $\theta \in \Theta$, $\mathbb{E}(|X_\theta|^p) < \infty$ où X_θ est une variable aléatoire de loi $\mathbb{P}_\theta$.
- Pour tout $j \in \{1, \cdots, p\}$, on estime dans un premier temps $\mathbb{E}(X^j)$ par l'estimateur empirique

$$\widehat{\mu}_j := \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

- Pour n assez grand, la valeur observée de $\widehat{\mu}_j$ est "proche" de $\mathbb{E}(X_{\theta_0}^j)$ (dont on connait l'expression en fonction de θ_0).
- On propose d'estimer θ_0 par la variable aléatoire $\widehat{\theta}_{0,n}$ solution du système d'équations d'inconnu θ :

$$\widehat{\mu}_j = \mathbb{E}(X_\theta^j),\ j \in \{1, \cdots, p\}.$$

- Si pour tout $\theta \in \Theta$, $\mathbb{P}_\theta$ est une loi discrète. La vraisemblance (observée) est la fonction

$$\theta \in \Theta \mapsto L(\theta; \mathbf{x}) := \prod_{i=1}^{n} \mathbb{P}[X_\theta = x_i],$$

où X_θ est une variable aléatoire de loi $\mathbb{P}_\theta$.

- Dans le cas où $\mathbb{P}_\theta$ est une loi absolument continue de densité $f(\cdot; \theta)$ pour tout $\theta \in \Theta$, la vraisemblance (observée) est la fonction

$$\theta \in \Theta \mapsto L(\theta; \mathbf{x}) := \prod_{i=1}^{n} f(x_i; \theta).$$

- La vraisemblance est la probabilité d'observer le vecteur $(x_1, \cdots, x_n)$ pour le vecteur aléatoire $(X_1, \cdots, X_n)$.
- L'idée de la méthode du maximum de vraisemblance est de trouver le paramètre $\theta \in \Theta$ le plus vraisemblable c'est-à-dire celui qui maximisera la probabilité d'observer les valeurs $(x_1, \cdots, x_n)$.
- L'estimateur du maximum de vraisemblance est la variable aléatoire

$$\widehat{\theta}_n := \underset{\theta \in \Theta}{\arg\max}\ L(\theta; \mathbf{X}).$$

- Les deux estimateurs précédents sont des estimateurs dit "ponctuels" c'est-à-dire que l'estimation est une valeur réelle.
- Il est souvent important de connaître la précision de cette estimation.
- Pour ce faire, on calculera l'intervalle de confiance $I_n(\alpha)$ de niveau $1-\alpha$ avec $\alpha \in]0,1[$ défini par

$$\mathbb{P}[I_n(\alpha) \ni \theta_0] = 1 - \alpha.$$

- La valeur de α est choisie par l'utilisateur (souvent $\alpha = 0.05$).

- $I_n(\alpha)$ est un intervalle aléatoire observable c'est-à-dire que $I_n(\alpha) := [L(\mathbf{X}), R(\mathbf{X})]$ où L et R sont des fonctions mesurables connues.
- La méthode la plus fréquemment utilisée pour construire un intervalle de confiance consiste à trouver une statistique $\varphi_n(\theta; \mathbf{X})$ telle que :
  1. la fonction $\theta \mapsto \varphi_n(\theta; \mathbf{X})$ est connue;
  2. la fonction $\theta \mapsto \varphi_n(\theta; \mathbf{X})$ est strictement monotone;
  3. la loi de $\varphi_n(\theta_0; \mathbf{X})$ est connue (et ne dépend donc pas de θ_0).
- Le point 2 assure que la fonction $\theta \mapsto \varphi_n(\theta; \mathbf{X})$ est inversible.
- Sans perte de généralité, on supposera que cette fonction est strictement décroissante.

- On note $z \mapsto \varphi_n^{-1}(z; \mathbf{X})$ l'inverse de $\varphi_n(\cdot; \mathbf{X})$.
- On note z_β le quantile d'ordre $\beta \in]0,1[$ de la loi de $\varphi_n(\theta_0; \mathbf{X})$ (i.e. $\mathbb{P}(\varphi_n(\theta_0; \mathbf{X}) \leq z_\beta) = \beta$).

$$\begin{aligned}
&\mathbb{P}\left(z_{\alpha/2} \leq \varphi_n(\theta_0; \mathbf{X}) \leq z_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(\varphi_n^{-1}(z_{1-\alpha/2}; \mathbf{X}) \leq \theta_0 \leq \varphi_n^{-1}(z_{\alpha/2}; \mathbf{X})\right) \\
&= 1 - \alpha.
\end{aligned}$$

L'intervalle de confiance de niveau $1-\alpha$ est donc

$$I_n(\alpha) = \left[\varphi_n^{-1}(z_{1-\alpha/2}; \mathbf{X}); \varphi_n^{-1}(z_{\alpha/2}; \mathbf{X})\right].$$

Exemple –

- $X_1, \cdots, X_n$ sont des variables aléatoires de loi normale de moyenne μ_0 inconnue et de variance $\sigma_0^2 > 0$ inconnue.
- La méthode des moments ou celle du maximum de vraisemblance conduisent aux mêmes estimateurs à savoir :

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2 .$$

- $\widehat{\mu}_n$ est un estimateur sans biais de μ_0.
- En revanche, l'estimateur sans biais de σ_0^2 est

$$\widehat{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2.$$

- Pour construire un intervalle de confiance pour μ_0, on utilise la statistique

$$\varphi_n(\mu; \mathbf{X}) := \frac{\sqrt{n}}{\widehat{s}_n^2} (\widehat{\mu}_n - \mu).$$

- Elle vérifie les points 1 et 2 données précédemment.
- On peut montrer que $\varphi_n(\mu_0; \mathbf{X})$ suit une loi de student à $n-1$ degrés de liberté.
- L'intervalle de confiance de niveau $1-\alpha$ pour μ_0 est donc

$$\left[\widehat{\mu}_n - t_{1-\alpha/2,n-1} \frac{\widehat{s}_n^2}{\sqrt{n}}; \widehat{\mu}_n - t_{\alpha/2,n-1} \frac{\widehat{s}_n^2}{\sqrt{n}} \right],$$

où $t_{\beta,n-1}$ est le quantile d'ordre $\beta \in]0,1[$ d'une loi de student à $n-1$ degrés de liberté.

- On peut remarquer que $t_{\alpha/2,n-1} = -t_{1-\alpha/2,n-1}$.

- On dispose des observations $\mathbf{x} = (x_1, \cdots, x_n)$ du vecteur aléatoire $\mathbf{X} = (X_1, \cdots, X_n)$ où $X_1, \cdots, X_n$ sont des variables aléatoires indépendantes de même loi $\mathbb{P}_X$.
- À partir des observations, on souhaite ici "choisir" (avec une probabilité d'erreur fixée) entre deux hypothèses :

i) $\mathbb{P}_X \in H_0$ où H_0 est appelée l'hypothèse nulle;

ii) $\mathbb{P}_X \in H_1$ où H_1 est l'hypothèse alternative.

- Exemple : pour une famille paramétrique de lois $\{\mathbb{P}_\theta;\ \theta \in \Theta\}$, on veut choisir entre $\mathbb{P}_X \in H_0 = \{\mathbb{P}_{\theta_0}\}$ contre $\mathbb{P}_X \in H_1 = \{\mathbb{P}_\theta;\ \theta \in \Theta \setminus \{\theta_0\}\}$.

On effectue le test de $H_0 \ :\ \theta = \theta_0$ contre $H_1 \ :\ \theta \neq \theta_0$.

- Un test est construit de la façon suivante. On se donne une statistique $S_n(\mathbf{X})$ telle que
  1. $S_n(\cdot)$ est une fonction mesurable **connue** à valeurs dans $\mathbb{R}$;
  2. si $\mathbb{P}_X \in H_0$, la loi $\mathbb{P}_0$ de $S_n(\mathbf{X})$ est **connue**.
- Pour une erreur α fixée par l'utilisateur, construire un test de $\mathbb{P}_X \in H_0$ contre $\mathbb{P}_X \in H_1$ revient à trouver une région $\mathcal{R}_\alpha \subset \mathbb{R}$ telle que $\mathbb{P}_0(S_n(\mathbf{X}) \in \mathcal{R}_\alpha) = 1 - \alpha$.
- Stratégie du test

i) Si $S_n(\mathbf{x}) \in \mathcal{R}_\alpha$, on ne rejette pas l'hypothèse nulle. Autrement dit, avec les observations dont on dispose, il est "statistiquement" possible que l'hypothèse nulle soit vraie (mais on n'en est pas sûr évidemment).

ii) Si $S_n(\mathbf{x}) \notin \mathcal{R}_\alpha$, on rejette l'hypothèse nulle au profit de l'hypothèse alternative.

Définition

La probabilité $\alpha = \mathbb{P}_0(S_n(\mathbf{X}) \notin \mathcal{R}_\alpha)$ est appelée l'erreur de type I. C'est la probabilité de rejeter l'hypothèse nulle alors qu'elle était vraie. Cette erreur est fixée par l'utilisateur (très souvent, on prend $\alpha = 0.05$).

- Lorsque $\mathcal{R}_\alpha$ est un intervalle de la forme $(-\infty, a]$, les logiciels de statistique donnent le résultat sous la forme d'une p-valeur.

Définition

Lorsque $\mathcal{R}_\alpha$ est un intervalle de la forme $(-\infty, a]$, la p-valeur est la probabilité définie par

$$p_{val} := \mathbb{P}_0\left(S_n(\mathbf{X}) \geq S_n(\mathbf{x})\right).$$

•Une valeur $p_{val} < \alpha$ conduira à rejeter H_0 avec une erreur de type I de α. La p-valeur est plus informative que le fait de dire simplement si on rejette ou non H_0. Plus la p-valeur est proche de 0 plus on est confiant dans le rejet de H_0.

Définition

L'erreur de type II notée β est la probabilité de ne pas rejeter H_0 alors que $\mathbb{P}_X \in H_1$.

• Cette erreur n'est pas contrôlée par l'utilisateur et elle est souvent difficile à calculer.

- Soient $X_1, \cdots, X_n$ des variables aléatoires indépendantes et de même loi $\mathbb{P}_X$ qui est une loi normale de moyenne et de variance inconnue.
- Pour une moyenne donnée $\mu_0 \in \mathbb{R}$, on souhaite faire un choix entre les hypothèses $\mathbb{P}_X \in H_0$ contre $\mathbb{P}_X \in H_1$ avec

$$H_0 = \{\mathcal{N}(\mu_0, \sigma^2);\ \sigma^2 > 0\} \text{ et } H_1 = \{\mathcal{N}(\mu, \sigma^2);\ \mu \neq \mu_0,\ \sigma^2 > 0\}.$$

- On écrira souvent les hypothèses sous la forme

$$H_0 \ : \ \mu = \mu_0 \text{ contre } H_1 \ : \ \mu \neq \mu_0.$$

- La statistique du test est

$$S_n(\mathbf{X}) := \frac{\sqrt{n}}{\widehat{s}_n^2} \left(\widehat{\mu}_n - \mu_0\right),$$

où l'on a utilisé les notations du paragraphe précédent.

- Comme μ_0 est connue (sa valeur est fixée par l'utilisateur), la fonction $S_n(\cdot)$ est connue.
- De plus, si $\mathbb{P}_X \in H_0$ (i.e., si $X_1, \cdots, X_n$ sont des variables aléatoires indépendantes de loi normale de moyenne μ_0 et de variance inconnue) alors $S_n(\mathbf{X})$ suit une loi (notée $\mathbb{P}_0$) de student à $n-1$ degrés de liberté (t_{n-1}).

Stratégie du test de student sur la moyenne d'une loi normale

- Une région d'acceptation possible est l'intervalle

$$\mathcal{R}_\alpha = [t_{\alpha/2,n-1}, t_{1-\alpha/2,n-1}]$$

où $t_{u,\nu}$ est le quantile d'ordre $u \in]0,1[$ d'une loi de student à ν degrés de liberté.

- Soient $\mathbf{x} = (x_1, \cdots, x_n)$ les observations dont on dispose.
- Pour une erreur de type I égale à α, on va rejeter H_0 si $S_n(\mathbf{x}) \notin [t_{\alpha/2,n-1}, t_{1-\alpha/2,n-1}]$ ou de manière équivalente si $|S_n(\mathbf{x})| > t_{1-\alpha/2,n-1}$.

- Contexte – Soit $\mathbf{X} := \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ un ensemble de n vecteurs aléatoires indépendants et de même loi $\mathbb{P}_{(X,Y)}$ qu'un vecteur aléatoire (X, Y).
- Les variables aléatoires X et Y sont discrètes à valeurs dans (respectivement) $E_X := \{a_1, \cdots, a_J\}$ et $E_Y := \{b_1, \cdots, b_K\}$.
- On souhaite effectuer le test de $\mathbb{P}_{(X,Y)} \in H_0$ contre $\mathbb{P}_{(X,Y)} \in H_1$ avec

$$H_0 = \left\{\mathbb{P}_{(X,Y)};\ \mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y\right\}$$

et

$$H_1 = \left\{\mathbb{P}_{(X,Y)};\ \mathbb{P}_{(X,Y)} \neq \mathbb{P}_X \otimes \mathbb{P}_Y\right\}.$$

- Autrement dit, on souhaite tester si X et Y sont indépendantes.

Construction de la statistique de test

- Pour tout $j \in \{1, \cdots, J\}$ et $k \in \{1, \cdots, K\}$, on pose

$$n_{j,k}(\mathbf{X}) := \sum_{i=1}^{n} \mathbb{I}\{X_i = a_j\}\mathbb{I}\{Y_i = b_k\};$$

$$n_{j,\bullet}(\mathbf{X}) := \sum_{k=1}^{K} n_{j,k}(\mathbf{X}) \; ; \; n_{\bullet,k}(\mathbf{X}) := \sum_{j=1}^{J} n_{j,k}(\mathbf{X}).$$

et

$$p_{j,k} = \mathbb{P}([X = a_j] \cap [Y = b_k])$$

- Les variables aléatoires

$$\{n_{j,k}(\mathbf{X});\ j \in \{1, \cdots, J\},\ k \in \{1, \cdots, K\}\}$$

suivent une loi multinomiale de paramètres n et

$$\{p_{j,k};\ j \in \{1, \cdots, J\},\ k \in \{1, \cdots, K\}\}.$$

- On a en particulier que $\mathbb{E}(n_{j,k}(\mathbf{X})) = np_{j,k}$.
- Sous l'hypothèse H_0 les effectifs attendus sont donc

$$\tilde{n}_{j,k} = n\mathbb{P}([X = a_j])\mathbb{P}([Y = b_k]).$$

- Les probabilités $\mathbb{P}([X = a_j])$ et $\mathbb{P}([Y = b_k])$ sont inconnues.
- Il y a J+K-2 probabilités à estimer.
- On les estime par

$$\hat{p}_{j,\bullet} := \frac{n_{j,\bullet}(\mathbf{X})}{n} \text{ et } \hat{p}_{\bullet,k} := \frac{n_{\bullet,k}(\mathbf{X})}{n}.$$

- On estime les effectifs attendus par

$$\hat{n}_{j,k}(\mathbf{X}) = \frac{n_{j,\bullet}(\mathbf{X}) n_{\bullet,k}(\mathbf{X})}{n}.$$

- On dresse deux tableaux.

Tableau des effectifs observés

$Y \backslash X$	$\cdots$	a_j	$\cdots$	Total
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
b_k	$\cdots$	$n_{j,k}(\mathbf{X})$	$\cdots$	$n_{\bullet,k}(\mathbf{X})$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
Total	$\cdots$	$n_{j,\bullet}(\mathbf{X})$	$\cdots$	n

Tableau des effectifs attendus

$Y \backslash X$	$\cdots$	a_j	$\cdots$	Total
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
b_k	$\cdots$	$n_{j,\bullet}(\mathbf{X})n_{\bullet,k}(\mathbf{X})/n$	$\cdots$	$n_{\bullet,k}(\mathbf{X})$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
Total	$\cdots$	$n_{j,\bullet}(\mathbf{X})$	$\cdots$	n

- La statistique du test est donnée par

$$S_n(\mathbf{X}) := \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{[n_{j,k}(\mathbf{X}) - \hat{n}_{j,k}(\mathbf{X})]^2}{\hat{n}_{j,k}(\mathbf{X})}.$$

- Sous H_0 (i.e., si X et Y sont indépendantes), on s'attend à observer une valeur de la statistique "proche" de 0.
- Sous H_0 on montre que $S_n(\mathbf{X})$ suit une loi du khi-deux à $(J-1)(K-1)$ degrés de liberrté.

Stratégie du test d'indépendance du khi-deux

- Une région d'acceptation possible est

$$\mathcal{R}_\alpha = [0, \chi_{1-\alpha}((J-1)(K-1))]$$

où $\chi_u(\nu)$ est le quantile d'ordre $u \in]0,1[$ d'une loi du khi-deux à ν degrés de liberté.

- Soient $\mathbf{x} = (x_1, \cdots, x_n)$ les observations dont on dispose.
- Pour une erreur de type I égale à α, on va rejeter H_0 si $|S_n(\mathbf{x})| > \chi_{1-\alpha}((J-1)(K-1))$.

• Contexte – Soit $\mathbf{X} := \{X_1, \cdots, X_n\}$ un ensemble de n variables aléatoires indépendantes et de même loi $\mathbb{P}_X$.
• La variable aléatoire X est discrète à valeurs dans $E_X := \{a_1, \cdots, a_J\}$.
• On souhaite effectuer le test de $\mathbb{P}_X \in H_0$ contre $\mathbb{P}_X \in H_1$ avec

$$H_0 = \{\mathbb{P}_\theta;\ \theta \in \Theta\} \text{ et } H_1 = \overline{H_0}.$$

Par exemple H_0 peut être l'ensemble des lois de Poisson de paramètre $\lambda > 0$ inconnu.
• Autrement dit, on souhaite tester si la loi de X appartient à la famille de loi paramétrique H_0.

Construction de la statistique de test

- Pour tout $j \in \{1, \cdots, J\}$, on pose

$$n_j(\mathbf{X}) := \sum_{i=1}^{n} \mathbb{I}\{X_i = a_j\} \text{ et } p_j = \mathbb{P}([X = a_j]).$$

Ainsi,

$$\mathbb{E}\left(n_j(\mathbf{X})\right) = np_j.$$

Sous l'hypothèse H_0, les effectifs attendus sont donc

$$\tilde{n}_{j,k} = n\mathbb{P}_\theta(\{a_j\}).$$

- Si le paramètre $\theta \in \Theta \subset \mathbb{R}^p$ est inconnu, on l'estime par $\widehat{\theta}_n$ (en utilisant la méthode du maximum de vraisemblance de préférence).
- On estime les effectifs attendus par

$$\hat{n}_j(\mathbf{X}) = n\mathbb{P}_{\widehat{\theta}_n}(\{a_j\}).$$

- La statistique du test est donnée par

$$S_n(\mathbf{X}) := \sum_{j=1}^{J} \frac{[n_j(\mathbf{X}) - \hat{n}_j(\mathbf{X})]^2}{\hat{n}_j(\mathbf{X})}.$$

- Sous H_0 (i.e., si la loi de X appartient à la famille paramétrique $\{\mathbb{P}_\theta;\ \theta \in \Theta\}$), on s'attend à observer une valeur de la statistique "proche" de 0.
- Sous H_0 on montre que $S_n(\mathbf{X})$ suit une loi du khi-deux à $J - p - 1$ degrés de liberté.

Stratégie du test d'indépendance du khi-deux

- Une région d'acceptation possible est

$$\mathcal{R}_\alpha = [0, \chi_{1-\alpha}(J-p-1)]$$

où $\chi_u(\nu)$ est le quantile d'ordre $u \in]0,1[$ d'une loi du khi-deux à ν degrés de liberté.

- Soient $\mathbf{x} = (x_1, \cdots, x_n)$ les observations dont on dispose.
- Pour une erreur de type I égale à α, on va rejeter H_0 si $|S_n(\mathbf{x})| > \chi_{1-\alpha}(J-n-1)$.

Régression linéaire simple

• Objectif – On observe n réalisations d'une variable aléatoires Y dont la loi dépend d'une variable non aléatoire $x \in \mathbb{R}$. Le modèle théorique de la régression linéaire simple est

$$Y = ax + b + U,$$

où Y est appelée la variable à expliquer, x la variable explicative et U est une variable aléatoire non observable avec $\mathbb{E}(U) = 0$ modélisant l'erreur d'observation de Y.

• En pratique, on dispose des valeurs $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ où y_i est l'observation de $Y_i = ax_i + b + U_i$ où l'on supposera ici que les variables aléatoires $U_1, \cdots, U_n$ sont indépendantes de même loi que U. On supposera également que U suit une loi $\mathcal{N}(0, \sigma^2)$.

• La première étape pour l'étude du modèle de régression linéaire est l'estimation des paramètres a et b.
• La méthode la plus couramment utilisée consiste à minimiser l'estimateur empirique de l'espérance $\mathbb{E}[(Y - \mathbb{E}(Y))^2]$.

Définition

L'estimateur des moindres carrés du modèle de régression linéaire simple $Y = ax + b + U$ est

$$(\widehat{a}_n, \widehat{b}_n) := \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - ax_i - b)^2 .$$

- En dérivant par rapport à a et b et en annulant la dérivée, on trouve facilement l'expression de $\widehat{a}_n$ et $\widehat{b}_n$.
- Quelques notations :

$$c(Y,x) := \frac{1}{n}\sum_{i=1}^{n} x_i Y_i - \overline{x}\,\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i(x_i - \overline{x}),$$

avec

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \text{ et } \overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

$$v(x) := \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- On peut vérifier que

$$\widehat{a}_n = \frac{c(Y,x)}{v(x)} \text{ et } \widehat{b}_n = \overline{Y} - \widehat{a}_n \overline{x}$$

- Il faut bien avoir conscience que $\widehat{a}_n$ et $\widehat{b}_n$ sont des variables aléatoires puisque Y est aléatoire. On notera $\tilde{a}_n$ et $\tilde{b}_n$ les valeurs observées :

$$\tilde{a}_n = \frac{c(y,x)}{v(x)} \text{ et } \tilde{b}_n = \overline{y} - \tilde{a}_n \overline{x}$$

• Le résultat suivant permet de décomposer la variance de Y en deux composantes : la variance résiduelle et la variance expliquée.

Proposition (Décomposition de la variance / version population)

En posant $\widehat{Y}_n(x) = \widehat{a}_n x + \widehat{b}_n$, *on a, si* $Y = aX + b + U$,

$$\mathbb{E}[(Y - \mathbb{E}(Y))^2] = \mathbb{E}[(Y - \widehat{Y}_n(x))^2] + \mathbb{E}[(\widehat{Y}_n(x) - \mathbb{E}(Y))^2].$$

• La variable aléatoire $\widehat{Y}_n(x)$ est appelée la prévision de Y pour une valeur donnée x de la variable explicative.

• Le terme $\mathbb{E}[(Y - \widehat{Y}_n(x))^2]$ est la variance résiduelle, celle mesurant la variance de l'erreur d'estimation. On souhaite que cette variance soit la plus petite possible.

• Le terme $\mathbb{E}[(\widehat{Y}_n(x) - \mathbb{E}(Y))^2]$ est la variance expliquée par le modèle. On souhaite que cette variance soit la plus grande possible.

- Quelques notations :

  La version échantillon de la variance de Y (somme des carrés totale) est (à un facteur $1/n$ près)

$$\mathrm{SCT}(Y, x) := \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2.$$

  La version échantillon de la variance résiduelle de Y (somme des carrés résiduelle) est (à un facteur $1/n$ près)

$$\mathrm{SCR}(Y, x) := \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_n(x_i)\right)^2.$$

- Notations (suite) :

  La version échantillon de la variance expliquée de Y (somme des carrés expliquée) est (à un facteur $1/n$ près)

$$\mathrm{SCE}(Y,x) := \sum_{i=1}^{n} \left(\widehat{Y}_n(x_i) - \overline{Y}\right)^2 = n\frac{[c(Y,x)]^2}{v(x)}.$$

Proposition (Décomposition de la variance / version échantillon)

$$\mathrm{SCT}(Y,x) = \mathrm{SCR}(Y,x) + \mathrm{SCE}(Y,x).$$

Définition (Coefficient de détermination)

Le coefficient de détermination linéaire est

$$R^2 = \frac{SCE(Y,x)}{SCT(Y,x)}.$$

- Interprétation – Plus la variance résiduelle est proche de zéro, plus le R^2 est proche de 1. Ainsi, une valeur de R^2 proche de 1 indique que la liaison entre Y et x semble bien être linéaire.
- Attention – Une valeur de R^2 proche de 1 n'assure en rien que les erreurs $U_1, \cdots, U_n$ sont indépendantes et de même loi normale centrée.
- On remarque que $R^2 = c^2(Y,x)/[v(x)v(Y)] = r^2(x,Y)$ où $r(x,Y)$ est le coefficient de corrélation linéaire.

- On peut également estimer la variance de modèle $\mathrm{Var}(U) = \sigma^2$ par

$$\widehat{\sigma}_n^2 = \frac{1}{n-2}\mathrm{SCR}(Y,x) = \frac{1}{n-2}\left(\mathrm{SCT}(Y,x) - n\frac{[c(Y,x)]^2}{v(x)}\right).$$

- On peut montrer que $\widehat{\sigma}_n^2$ est un estimateur sans biais de σ^2.
- La valeur observée de $\widehat{\sigma}_n^2$ est

$$\tilde{\sigma}_n^2 = \frac{1}{n-2}\mathrm{SCR}(y,x) = \frac{1}{n-2}\left(\mathrm{SCT}(y,x) - n\frac{[c(y,x)]^2}{v(x)}\right).$$

- Sous l'hypothèse que $Y = ax + b + U$ où $U \sim \mathcal{N}(0, \sigma^2)$, on connait les lois des différents estimateurs.
  - Loi de $\widehat{a}_n$: on a $\mathbb{E}(\widehat{a}_n) = a$ et $\mathrm{Var}(\widehat{a}_n) = \sigma^2/(nv(x))$. De plus, en notant $\widehat{\sigma}^2_{a,n} = \widehat{\sigma}^2_n/(nv(x))$ la variance estimée de $\widehat{a}_n$,

$$\frac{\widehat{a}_n - a}{\widehat{\sigma}_{a,n}} \sim t_{n-2}.$$

On en déduit un intervalle de confiance au niveau $1-\alpha$ de a :

$$I_a(1-\alpha) := \left[\widehat{a}_n - \widehat{\sigma}_{a,n} t_{1-\alpha/2,n-2} \; ; \;\; \widehat{a}_n + \widehat{\sigma}_{a,n} t_{1-\alpha/2,n-2}\right].$$

Enfin, la statistique du test de $H_0 \;:\; a = a_0$ est

$$S_n(Y) = \frac{\widehat{a}_n - a_0}{\widehat{\sigma}_{a,n}}$$

Remarques –

- On peut vérifier facilement que rejeter l'hypothèse $H_0 \ : \ a = a_0$ au profit de $H_1 \ : \ a \neq a_0$ est équivalent à dire que $a_0 \notin \tilde{I}_a(1-\alpha)$ où $\tilde{I}_a(1-\alpha)$ est l'intervalle de confiance observé.
- Tester la significativité du modèle signifie tester $H_0 \ : a = 0$ contre $H_1 \ : \ a \neq 0$. On dira que le modèle est significatif lorsqu'on rejette H_0.
- L'intervalle de confiance $I_a(1-\alpha)$ est dit bilatéral. L'intervalle

$$\left[\widehat{a}_n - \widehat{\sigma}_{a,n} t_{1-\alpha/2,n-2} \ ; \ \ +\infty\right[,$$

est un intervalle unilatéral à gauche.

- L'intervalle

$$\left]-\infty \ ; \ \widehat{a}_n + \widehat{\sigma}_{a,n} t_{1-\alpha/2,n-2}\right] ,$$

est un intervalle unilatéral à droite.

- Loi de $\widehat{b}_n$: on a $\mathbb{E}(\widehat{b}_n) = b$ et

$$\mathrm{Var}(\widehat{b}_n) = \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{v(x)}\right).$$

De plus, en notant

$$\widehat{\sigma}^2_{b,n} = \frac{\widehat{\sigma}^2_n}{n}\left(1 + \frac{\overline{x}^2}{v(x)}\right),$$

la variance estimée de $\widehat{b}_n$, on a $(\widehat{b}_n - b)/\widehat{\sigma}_{b,n} \sim t_{n-2}$. De la même façon que précédemment, ce résultat nous permet de construire un intervalle de confiance pour b et de proposer une statistique de test de l'hypothèse nulle $H_0 \ : \ b = b_0$.

- Loi de $\widehat{\sigma}_n$: on a

$$\frac{(n-2)\widehat{\sigma}_n^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Ce résultat nous permet de construire un intervalle de confiance pour σ^2 et de proposer une statistique de test de l'hypothèse nulle $H_0 \ : \ \sigma^2 = \sigma_0^2$.

- Loi de la prévision : pour tout x_0, si on note $Y_{x_0} = ax_0 + b + U$, on a

$$\left(Y_{x_0} - \widehat{Y}_n(x_0)\right) \Bigg/ \left[\frac{\widehat{\sigma}_n^2}{n}\left(1 + n + \frac{(\overline{x} - x_0)^2}{v(x)}\right)\right]^{1/2} \sim t_{n-2}.$$

Un intervalle de confiance au niveau $1-\alpha$ de la prévision de Y associé à la valeur x_0 est

$$\left[\widehat{Y}_n(x_0) \pm \left[\frac{\widehat{\sigma}_n^2}{n}\left(1 + n + \frac{(\overline{x} - x_0)^2}{v(x)}\right)\right]^{1/2} t_{1-\alpha/2,n-2}\right].$$

En régression linéaire multiple, on souhaite expliquer la variable Y par plusieurs covariables $x^{(1)}, \ldots, x^{(p)}$. Le modèle s'écrit :

$$Y = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)} + U,$$

où $U \sim \mathcal{N}(0,1)$. En écriture vectorielle, on a $Y = \beta^\top \mathbf{x} + U$ avec $\beta = (\beta_1, \ldots, \beta_p)$ et $\mathbf{x} = (1, x^{(1)}, \ldots, x^{(n)})^\top$. En pratique, on observe les couples $\{(\mathbf{x}_i, y_i);\ i = 1, \ldots, n\}$, où y_i est la réalisation de $Y_i = \beta^\top \mathbf{x}_i + U_i$. Sous sa forme matricielle, le modèle s'écrit donc

$$\mathbf{Y} = \beta^\top \mathbf{X} + \mathbf{U},$$

avec $\mathbf{X}$ la matrice de dimension $n \times p + 1$ dont la $i$ème ligne est le vecteur $\mathbf{x}_i$, $\mathbf{Y} = (Y_1, \ldots, Y_n)$ et $\mathbf{U} = (U_1, \ldots, U_n)^\top$.

On estime le vecteur β par la méthode des moindres carrés de la façon suivante :

$$\widehat{\beta}_n := \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta^\top \mathbf{x}_i \right)^2$$

On peut montrer que

$$\widehat{\beta}_n = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Afin de pouvoir inverser la matrice $\mathbf{X}^\top \mathbf{X}$ il faut que les variables explicatives $x^{(1)}, \ldots, x^{(p)}$ ne soient pas corrélées.