Statistiques et Probabilités

UIT II Informatique - S3 -M3201

Maha Elbayad maha.elbayad@inria.fr
Franck Corset franck.corset@univ-grenoble-alpes.fr

Cours disponible sur **https://fcorset.github.io/cours/cours.html**

7 Octobre 2019

Estimation

Dans un problème statistique, la loi d'une v.a. observée est inconnue.

Les étapes de la statistique inférentielle :

1. Observation d'une variable X sur un échantillon construit d'une façon aléatoire et indépendante dans la population totale i.e $X_1, X_2, \dots X_N \sim X$ i.i.d.
2. Chaque X_i a pour réalisation x_i. On fait une étude descriptive de $x_1, \dots, x_n$ (histogramme, moyenne, ...).
3. Au vu de l'étude descriptive, trouver une loi de probabilité 'acceptable' pour les variables $X_1, \dots, X_n$ et estimer ses paramètres (par exemple: μ et σ d'une gaussienne).

Estimateurs

Estimateur

Pour un paramètre inconnu θ, un estimateur est une fonction des données $(x_1, \ldots, x_n)$ qui permet d'approximer θ.

- Avant d'avoir observé les réalisations $(x_1, \ldots, x_n)$, l'estimateur est une variable aléatoire.
- Une fois les données collectées, l'estimation est la valeur de l'estimateur pour ces données.

Estimateurs

Définition formelle

Un estimateur de θ est une v.a., fonction d'un échantillon à n éléments de X:

$$\hat{\theta}_n = f(X_1, \ldots, X_n)$$

dont les valeurs observées, dites estimations, seront probablement "suffisamment proches" de la valeur inconnue θ.

Qualité des estimateurs

- Le biais de $\hat{\theta}_n$ est la différence entre l'espérance de $\hat{\theta}_n$ et la vraie valeur θ.

$$\text{Biais}(\hat{\theta}_n) = \mathbb{E}\left[\hat{\theta}_n\right] - \theta$$

- L'erreur quadratique est l'espérance des carrés des différences:

$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right] = \text{Bias}(\hat{\theta}_n)^2 + \text{Var}[\theta_n]$$

MSE (ou enocre EQM) permet de classer les estimatuers, **plus faible est MSE meilleur est l'estimateur.**
Souvent on note un estimateur d'une valeur θ avec T (laissant tomber l'indice n):

$$\text{Biais}(\mathrm{T}) = \mathbb{E}[\mathrm{T}] - \theta$$

$$\text{EQM}(\mathrm{T}) = \text{MSE}(\mathrm{T}) = \mathbb{E}\left[(\mathrm{T} - \theta)^2\right] = (\mathbb{E}[\mathrm{T}] - \theta)^2 + \text{Var}[\mathrm{T}] = \text{Biais}(\mathrm{T})^2 + \text{Var}[\mathrm{T}]$$

Qualité des estimateurs

Propriétés

Un estimateur $\hat{\theta}_n$ est:

- **Sans biais** si Biais$(\hat{\theta}_n) = 0$ i.e. les valeurs de $\hat{\theta}_n$ sont centrées autour de θ.
- **Asymptotiquement sans biais** si $Biais(\hat{\theta}_n) \xrightarrow[n \to +\infty]{} 0$
- **Consistant**

$$\forall \epsilon > 0, \lim_{n \to +\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

Quand la taille de l'echantillon augmente, la probabilité de s'éloigner de θ de plus d'un ϵ (petit) tend vers 0.

Exemples d'estimateurs

1. La fréquence empirique d'un événement:

$$\frac{\text{nombre d'occurences d'un événement}}{\text{nombre total d'observations}}$$

est un estimateur **sans biais** consistant de la probabilité de cet évènement.

2. La moyenne empirique d'un échantillon est un estimateur **sans biais** consistant de l'espérance théorique θ de ces variables:

$$\hat{\theta}_n(X_1, \ldots, X_n) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Par la linéarité de $\mathbb{E}$:

$$\text{Biais}(\hat{\theta}_n) = \mathbb{E}\left[\hat{\theta}_n\right] - \mathbb{E}[X] = \frac{1}{n}\sum_{i} \mathbb{E}[X] - \mathbb{E}[X] = 0$$

Exemples d'estimateurs

4. La variance empirique notée S_n^2 d'un échantillon (lorsque la moyenne est inconnue) est un estimateur de la variance théorique σ^2:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Cet estimateur est **biaisé** avec $\text{Biais}(S_n^2) = \frac{n-1}{n}\sigma^2$.
Ainsi, on obtient un estimateur **sans biais** en multipliant la variance empirique par $\frac{n}{n-1}$ que l'on note ${S'_n}^2$:

$${S'_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

C'est cette dernière quantité qu'on obtient avec la fonction `var()` de `R`.

Exemples d'estimateurs

Preuve: S_n^2 est **biaisé**.

$$nS_n^2 = \sum_i \left(X_i^2 + \bar{X}^2 - 2\bar{X}X_i\right) = \sum_i X_i^2 + n\bar{X}^2 - 2\bar{X}\sum_i X_i$$
$$= \sum_i X_i^2 + n\bar{X}^2 - 2(n\bar{X}\bar{X}) = \sum_i X_i^2 - n\bar{X}^2$$

On sait que $\forall i, \mathbb{E}\left[X_i^2\right] - \mathbb{E}[X_i]^2 = \sigma^2$, et que $\forall i \neq j$, $\mathbb{E}[X_iX_j] = \mathbb{E}[X_i]^2$ d'où

$$\mathbb{E}\left[\bar{X}^2\right] = \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j} X_iX_j\right] = \frac{1}{n^2}\left(\sum_{i\neq j}\mathbb{E}[X_i]^2 + \sum_i \mathbb{E}\left[X_i^2\right]\right) = \frac{1}{n^2}\left(n(n-1)\mathbb{E}[X_i]^2 + n(\sigma^2 + \mathbb{E}[X_i]^2)\right)$$
$$= \mathbb{E}[X_i]^2 + \frac{\sigma^2}{n}$$

Dóù:

$$n\mathbb{E}\left[S_n^2\right] = n\mathbb{E}\left[X_i^2\right] - n\mathbb{E}\left[\bar{X}^2\right] = n(\sigma^2 + \mathbb{E}[X_i]^2) - n\mathbb{E}[X_i]^2 - \sigma^2 = (n-1)\sigma^2$$

Intervalle de confiance

Le rôle des intervalles de confiance est de donner une idée sur la **précision de l'estimation** d'un paramètre θ avec des estimateurs.

Problème:
Pour $\alpha \in]0,1[$, trouver deux statistiques T_1 et T_2 telles que $p(T_1 \leq \theta \leq T_2) = 1 - \alpha$

- L'intervalle $[T_1, T_2]$ est un intervalle aléatoire appelé intervalle de confiance.
- α est le risque d'erreur.

  $$P(\theta \notin [T_1, T_2]) = \alpha$$

  On se trompe en moyenne 100α fois sur 100.
- $(1 - \alpha)$ est appelé niveau de confiance ou coefficient de sécurité

  $$P(\theta \in [T_1, T_2]) = \alpha$$

  .

Intervalle de confiance pour une moyenne

$X \sim \mathcal{N}(\mu, \sigma^2)$

On suppose que X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. On rappelle que la moyenne empirique et que la variance empirique sont données par

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- Si σ^2 est connue, un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ est

$$\left[\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$ i.e.
$P(X \leq u_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.
(Sur R: $u_{1-\alpha/2}$ =qnorm$(1-\alpha/2)$)

Intervalle de confiance pour une moyenne

$X \sim \mathcal{N}(\mu, \sigma^2)$

On suppose que X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. On rappelle que la moyenne empirique et que la variance empirique sont données par

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- Si σ^2 est inconnue, un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ est

$$\left[\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student de paramètre $n-1$.
(Sur R: $t_{1-\alpha/2} =$ `qt(1-α/2, df=n-1)`).

Intervalle de confiance pour une moyenne

Echantillon de grande taille

Pour de grands échantillons, sans hypothèse de normalité, un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ est

$$\left[\bar{X} - u_{1-\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + u_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

Intervalle de confiance pour une variance

$X \sim \mathcal{N}(\mu, \sigma^2)$

On se place dans le cas où X suit une loi normale, $\mathcal{N}(\mu, \sigma^2)$.
Un intervalle de confiance de niveau $1-\alpha$ pour la variance σ^2 est

$$\left[\frac{nS^2}{q^{n-1}_{1-\alpha/2}}; \frac{nS^2}{q^{n-1}_{\alpha/2}}\right] = \left[\frac{(n-1)(S')^2}{q^{n-1}_{1-\alpha/2}}; \frac{(n-1)(S')^2}{q^{n-1}_{\alpha/2}}\right]$$

où $q^{n-1}_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi du chi-2 de paramètre $n-1$ et $q^{n-1}_{\alpha/2}$ son quantile d'ordre $\alpha/2$.

Intervalle de confiance pour une proportion

Echantillon de grande taille

On suppose que l'on est en présence d'un échantillon de grande taille (en pratique $n \geq 30$). Un intervalle de confiance de niveau $(1-\alpha)$ pour une proportion p inconnue est

$$\left[\hat{p} - u_{1-\alpha/2}\sqrt{(\frac{\hat{p}(1-\hat{p})}{n})}; \hat{p} + u_{1-\alpha/2}\sqrt{(\frac{\hat{p}(1-\hat{p})}{n})}\right].$$

où n est la taille de l'échantillon, $\hat{p}$ la fréquence empirique et $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.