

Statistiques, séance n°3

EPITA

Avril 2020

Objectif de ce cours

- L'estimation
- Cf. illustrations en R

Problématique

- A partir d'un échantillon, retrouver les caractéristiques permettant de le modéliser
- Exemple : Montrer qu'une pièce, pour un jeu de pile ou face n'est pas « truquée »
 - On joue de nombreuses fois
 - On calcul la fréquence d'apparition des piles et faces
 - On essaie de conclure
- Expérience avec R
- La fréquence observée est une « **statistique** »
- Cette statistique est un « estimateur » de la « vraie fréquence »
- C'est-à-dire du paramètre **p** permettant de simuler la loi de **Bernouilli correspondante**

Statistiques et estimateurs

- Échantillon $x_1, x_2, \dots, x_n$ issu des variables aléatoires $X_1, X_2, \dots, X_n$
- Une statistique T est une variable aléatoire fonction des $X_1, X_2, \dots, X_n$
- Pour un paramètre θ , estimateur T de θ
- Exemples :
 - Espérance m estimateur : $\bar{X} = (X_1 + \dots + X_n)/n$
 - Variance σ^2 estimateur : $S^2 = \frac{1}{n}((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2)$
 - Probabilité d'un événement p : sa fréquence empirique F

Conséquence de la loi des grands nombres

- *Presque sûrement :*

- $\bar{X} \rightarrow m$

- $S^2 \rightarrow \sigma^2$

- $F \rightarrow p$

Qualité d'un estimateur

- **Convergence**

- $T \rightarrow \theta$

- **Précision**

$$T - \theta = T - \mathbb{E}(T) + \mathbb{E}(T) - \theta$$

- **Biais**

$$\mathbb{E}(T) - \theta$$

- **Erreur quadratique moyenne**

$$\begin{aligned} & \mathbb{E} \left((T - \theta)^2 \right) \\ & \mathbb{E} \left((T - \theta)^2 \right) = V(T) + (\mathbb{E}(T) - \theta)^2 \end{aligned}$$

- ... ??? **Estimateur sans biais de variance minimale**

Précaution

- Pour la variance,
- L' estimateur :

$$S^2 = \frac{1}{n} \left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right) \text{ est biaisé}$$

$$\mathbb{E}(S^2) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

- On préfère l'estimateur sans biais :

$$S^{*2} = \frac{1}{n-1} \left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right)$$

$$\mathbb{E}(S^{*2}) = \sigma^2$$

- L'existence d'estimateurs sans biais de variance minimale est réservée à un autre cours
- Dans ce cours, le problème posé est 'Etant donné un paramètre θ et un estimateur T de ce paramètre, étant donné un échantillon $x_1, x_2, \dots, x_n$ proposer un encadrement de θ dans un intervalle $[\theta_1, \theta_2]$ avec une « faible » probabilité d'erreur'
- On montre la méthode sur des exemples

Echantillon représentatif

- Tirage permettant à tout individu d'une population de figurer dans l'échantillon avec la même probabilité

Pour le slide suivant

- Une V.A. qui m'intéresse : X
- Un échantillon représentation de la population ...
- On va estimer : $\mu = \mathbb{E}(X)$
- Comment donner cette estimation ???

Estimation d'une moyenne lorsque la variance est connue

- On se fixe une marge d'erreur α (ex : 5%)
- L'estimateur de l'espérance μ est $\overline{X}_n = \frac{1}{n} \sum_i X_i$
- Il faut d'abord connaître le comportement de l'estimateur
 - Le théorème de la limite centrale permet, lorsque n est grand, d'approcher $\overline{X}_n$ par une loi $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
 - Par symétrie de la loi normale, on choisit $u_{\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$; dans le cas où $\alpha = 5\%$, $u_{\frac{\alpha}{2}} = 1,96$
 - Alors $\mathbb{P}\left(\left|\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu)\right| > u_{\frac{\alpha}{2}}\right) = \alpha$ $-u_{\frac{\alpha}{2}} < \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) < u_{\frac{\alpha}{2}}$
 - Avec une probabilité $1 - \alpha$, l'intervalle de confiance est, pour l'échantillon : $x_1, \dots, x_n$
$$\overline{x}_n - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{x}_n + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Largeur de l'intervalle de confiance

$$2u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Estimation d'une proportion

- On se fixe une marge d'erreur α (ex : 5%)
- Cette fois les X_i suivent une loi de Bernoulli de paramètre p inconnu, avec une variance $p(1 - p)$ elle aussi inconnue
- En approximant la fréquence (variable aléatoire) F par une loi normale, soit f la fréquence observée sur l'échantillon, avec une probabilité $1 - \alpha$:

$$f - u_{\frac{\alpha}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}} < p < f + u_{\frac{\alpha}{2}} \frac{\sqrt{p(p-1)}}{\sqrt{n}}$$

- Comme p est inconnu, plusieurs possibilité : résoudre une équation du 2nd degré, quand n est grand approximer $\sqrt{p(p-1)}$ par $\sqrt{f(f-1)}$, sinon remarquer qu'on a toujours : $\sqrt{p(p-1)} < \frac{1}{2}$. Cette dernière inégalité donne :

$$f - u_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}} < p < f + u_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}$$

Intervalle de confiance pour l'espérance d'une variable normale, quand la variance σ^2 est inconnue

- La variable aléatoire $T = \frac{\sqrt{n-1}}{S_n} (\overline{X}_n - \mu)$ suit une loi de Student à $n - 1$ degrés de libertés.
- En R, la fonction de répartition de la loi de Student est qt. On prend $t_{\alpha/2} \leftarrow qt\left(1 - \frac{\alpha}{2}, n - 1\right)$ *convention du langage R*

- On obtient (même raisonnement) :

$$\overline{x}_n - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}} < \mu < \overline{x}_n + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}}$$

- En pratique, lorsque n est petit $t_{\frac{\alpha}{2}}$ est nettement plus élevé que $u_{\frac{\alpha}{2}}$!
- Le fait de ne pas connaître la variance augmente la largeur de l'intervalle de confiance.