M1 Mathmématiques et Intelligence artificielle - Paris-Saclay
Apprentissage statistique, 2022-2023
christine.keribin@universite-paris-saclay.fr

# Semaine 1: Exercices de révision

Les exercices 1 à 3 portent sur les estimateurs, l'exercice 4 sur un exemple d'utilisation de la méthode des moments, les exercices 5 à 7 sur les tests et ICs.

**Exercice 1.**
Soit un $n$-échantillon $(X_1, \dots, X_n)$ iid de loi de Bernoulli $\mathcal{B}(1, p)$

1. Soit $\widehat{p} = \sum_{i=1}^n X_i/n$. $\widehat{p}$ est-il un estimateur de $p$ biaisé? consistant?
2. Soit $S_n = \sum_i X_i$. Déterminer la loi de $S_n$.

   Soit $g(S_n)$ un autre estimateur sans biais de $p$ fonction de $S_n$. Montrer que pour tout $x \in \mathbb{R}^+$
   $$\sum_{k=0}^{n} \left( g(k) - \frac{k}{n} \right) \binom{n}{k} x^k = 0$$
   En déduire que $\widehat{p}$ est le seul estimateur sans biais de $p$ fonction de $S_n$.
3. Construire un estimateur de la variance de $\bar{X}$ en utilisant uniquement le premier moment, puis en utilisant également le moment d'ordre 2. Montrer que les deux estimateurs coïncident.
4. Montrer qu'il n'y a pas unicité de l'estimateur des moments en considérant le cas de l'estimation du paramètre $\lambda$ de la loi exponentielle (de densité $f(x) = \lambda^{-1} \exp(-x/\lambda)$ sur $\mathbb{R}^+$) et ses premier et deuxième moments.

**Exercice 2.**
Soit un $n$-échantillon iid $(X_1, \dots, X_n)$ de loi gaussienne d'espérance $\mu$ et de variance $\sigma^2 > 0$, toutes les deux inconnues. On s'intéresse à l'estimation de $\sigma^2$.

Rappel: $K_{n-1} = \sum_i (X_i - \bar{X})^2/\sigma^2 \sim \chi^2(n-1)$. La densité de la loi du Khi-deux à $k$ degrés de liberté $\chi^2(k)$ est définie sur $\mathbb{R}^+$ par

$$f(x) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} \exp(-x/2)$$

1. Déterminer un estimateur $V_n$ par la méthode des moments.

   Quelle est son espérance?
2. Calculer son risque quadratique.
3. Proposer un estimateur $\widehat{\sigma}^2$ sans biais. Comparer son risque avec celui de l'estimateur empirique.
4. Déterminer $a$ pour que $T_a = a \sum_{i=1}^n (X_i - \bar{X})^2$ soit le meilleur estimateur parmi les estimateurs de cette forme.
5. Vrai ou Faux?
   (a) Un estimateur de risque minimum est forcément de variance minimum

(b) Un estimateur non biaisé est de risque minimum

(c) Un estimateur dont la variance tend vers 0 est consistant

6. L'estimateur $\sqrt{\widehat{\sigma}^2}$ est-il biaisé pour estimer $\sigma$? Commenter.

**Exercice 3.**
Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un $n$-échantillon iid de loi mère telle que $\mathbb{E}(X_1^2)$ et $\mathbb{E}(Y_1^2)$ soient finis. Soit

$$C_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

1. Montrer que $C_n = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}$.
2. Montrer que $C_n$ est biaisé pour estimer $\mathrm{cov}(X_1, Y_1)$.
3. Montrer que $C_n$ est un estimateur consistant de $\mathrm{cov}(X_1, Y_1)$.
4. Proposer un estimateur consistant et sans biais de $\mathrm{cov}(X_1, Y_1)$.

**Exercice 4.**

On considère un $n$-échantillon de loi beta de densité définie sur $[0; 1]$ par

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

où $\alpha > 0$ et $\beta > 0$ On rappelle que

$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta} \text{ et } \mathrm{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Construire un estimateur des moments de $(\alpha, \beta)$. Est-il consistant?

**Exercice 5.**
Soit $(X_1, ..., X_{25})$ un échantillon de loi gaussienne d'espérance $\mu$ inconnue et de variance $v = 100$ connue.

1. Construire un test de niveau $\alpha = 0.10$ de l'hypothèse nulle "$\mu = 0$" contre l'hypothèse alternative "$\mu = 1.5$", fondé sur la moyenne empirique, estimateur du paramètre $\mu$.

   On observe $\overline{x} = 1$. Quelle est la décision du test?

   Quelle est l'erreur de seconde espèce du test?
2. Répondre à la question précédente si $v = 9$.
3. Comment modifier le test si l'alternative est "$\mu = -1.5$"?
4. On souhaite tester $(H_0) : \mu = 2$ contre $(H_1) : \mu < 2$. Définir la région de rejet. Calculer la puissance du test et étudier ses variations en fonction de $\mu$, $n$ et $\sigma$.

**Exercice 6.**
On dispose de 100 mesures de la vitesse de la lumière (km $s^{-1}$) effectuées en 1879 par Michelson, lors de la célèbre expérience qui lui valut le prix Nobel de physique en 1907[1]. Avec les instruments de cette époque, les scientifiques avaient calculé une vitesse de la lumière avec une légère imprécision: $c_{1879} = 299840$ km/s. Aujourd'hui, la vitesse de la lumière est établie à $c_{2018} = 299792$ km/s.

Dans tout l'exercice, les niveaux des tests sont pris à 5%. On modélisera les mesures de Michelson comme des variables aléatoires iid gaussiennes d'espérance et de variance inconnues.

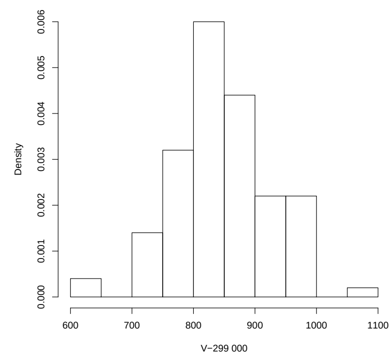


Figure 1: Histogramme des mesures de Michelson

1. On souhaite tester l'hypothèse nulle que les mesures de Michelson sont conformes à la valeur de la vitesse de la lumière communément admise en 1879.

   Construire un test approprié.

   Sur l'échantillon de Michelson, la moyenne observée vaut $m = 299852$ et l'écart-type corrigé vaut $s = 79$. Quelle décision prendre et à quel risque?

   Calculer la p-value et vérifier cette décision.

2. Construire un intervalle de confiance bilatère de niveau 95% de la vitesse de la lumière à partir de l'échantillon de Michelson.

3. Les observations de Michelson sont-elles en accord avec la valeur de la vitesse de la lumière mesurée aujourd'hui? Justifier la réponse par un test.

4. On souhaite tester la précision $\sigma$ de la mesure. Définir la statistique de test, la région de rejet, la p-value, la décision et son risque dans les deux cas suivants:

   (a) $(H_0) : \sigma = 90$ contre $(H_1) : \sigma < 90$.

   (b) L'écart-type $\sigma$ de la mesure peut-il être considéré comme égal à 91 ? Commenter.

[1] John Rice: Mathematical Statistics and Data Analysis, Duxbury Press 1995

**Exercice 7.**
Un fournisseur de pièces mécaniques indique que son taux de pièces défectueuses ne dépasse pas 5%. Pour tester cette affirmation, le service qualité prélève au hasard 200 pièces dans un lot de 10000 pièces et trouve 14 pièces défectueuses.
Peut-on affirmer, avec une faible erreur (5%) que le lot ne remplit pas la condition de qualité annoncée?
Calculer la p-value du test.
Construire un intervalle de confiance bilatéral de niveau 95%.