ENSTA ParisTech 1ère année
Statistiques (MA101), 2017-2018
christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

TD3: Echantillons gaussiens

Exercice 1.
Soit un n-échantillon de loi gaussienne d'espérance μ et de variance σ^2, toutes les deux inconnues. On s'intéresse à l'estimation de σ^2.

1. Déterminer l'estimateur empirique par la méthode des moments et calculer son biais.
2. Calculer l'estimateur du maximum de vraisemblance S_n^2. Quelle est sa loi?
3. Calculer le risque de S_n^2. Est-il consistant?
4. Proposer un estimateur sans biais $\widehat{\sigma}^2$. Comparer son risque avec celui de l'estimateur empirique.
5. Déterminer le comportement asymptotique. Comment ce résultat est-il modifié si on fait uniquement l'hypothèse d'une loi de carré intégrable?

Correction. *1. On utilise les deux premiers moments:* $m_1 = \mathbb{E}(X) = \mu$ *et* $m_2 = \mathbb{E}(X^2) = \sigma^2 + \mu^2$ *qu'on estime respectivement par*

$$\widehat{m}_1 = \frac{1}{n}\sum_{i^1}^{n} X_i; \quad \widehat{m}_2 = \frac{1}{n}\sum_{i^1}^{n} X_i^2$$

La variance de la loi de X_1 *est définie par* $\sigma^2 = \mathbb{E}[(X_1 - \mathbb{E}(X))^2] = \mathbb{E}(X_1^2) - (\mathbb{E}(X))^2 = m_2 - m_1^2$. *D'où*

$$V_n = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

On en déduit l'espérance de V_n

$$\begin{aligned}
\mathbb{E}(V_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i - \mu + \mu - \bar{X})^2] \\
&= \frac{1}{n}\sum_{i}\mathbb{E}\left((X_i-\mu)^2\right) + \frac{2}{n}\mathbb{E}\left(\sum_i (X_i-\mu)(\mu-\bar{X})\right) + \frac{n}{n}\mathbb{E}\left((\mu-\bar{X})^2\right) \\
&= \frac{n}{n}\sigma^2 - 2\,\mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{X}) = \frac{n}{n}\sigma^2 - \mathrm{Var}(\bar{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}$$

Ce calcul a été fait pour tout type de loi de carré intégrable et n'est pas spécifique à la loi gaussienne.

2. *Le paramètre est* $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$. *La log-vraisemblance des observations s'écrit:*

$$\log L(\theta; X) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}$$

L'EMV est solution des équations normales

$$\frac{\partial}{\partial \mu}\log L(\theta; X) = \sum_i (X_i - \mu)/\sigma^2; \quad \frac{\partial}{\partial \sigma^2}\log L(\theta; X) = -\frac{n}{2\sigma^2} + \frac{\sum_i (X_i - \mu)^2}{2\sigma^4}$$

d'où $\widehat{\mu}_{EMV} = \bar{X}$, $\widehat{\sigma}^2_{EMV} = V_n$, *identiques à l'estimateur des moments. On vérifie que le hessien du système est défini négatif. Dans le cas gaussien,* $\widehat{\mu}_{EMV}$ *suit une loi (exacte)* $\mathcal{N}(\mu, \sigma^2/n)$ *et* $nV_n/\sigma^2 = \sum_{i=1}^n (X_i - \bar{X})^2/\sigma^2 = K_{n-1}$ *une loi du Khi-deux à* $n-1$ *degrés de liberté*

3. *En utilisant le fait qu'une loi* $K_n \sim \chi^2(n)$ *est d'espérance* n *et de variance* $2n$*:*

$$\mathrm{Var}(V_n) = \mathrm{Var}\left(\frac{\sigma^2}{n} K_{n-1}\right) = 2(n-1)\frac{\sigma^4}{n^2}$$

$$\begin{aligned} R(V_n) &= \mathrm{Var}(V_n) + (\mathbb{E}(V_n) - \sigma^2)^2 \\ &= 2(n-1)\frac{\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 \\ &= 2(n-1)\frac{\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = (2n-1)\frac{\sigma^4}{n^2} \end{aligned}$$

Le risque tendant vers 0 quand n *tend vers l'infini, l'estimateur est convergent en moyenne quadratique, donc consistant.*

4. *L'estimateur* $\widehat{\sigma}^2_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2) = \sigma^2 K_{n-1}/(n-1)$ *est sans biais* $\mathbb{E}(\sigma^2_n) = \sigma^2$ *et de variance égale à son risque*

$$\mathrm{Var}(\widehat{\sigma}^2_n) = \frac{2\sigma^4}{n-1} = R(\widehat{\sigma}^2_n)$$

On en déduit que V_n *domine* $\widehat{\sigma}^2_n$*:*

$$\begin{aligned} R(\widehat{\sigma}^2_n) - R(V_n) &= \frac{2\sigma^4}{n-1} - (2n-1)\frac{\sigma^4}{n^2} \\ &= \frac{\sigma^4}{n^2(n-1)}(2n^2 - (n-1)(2n-1)) \\ &= \frac{3n-1}{(n-1)n^2}\sigma^4 > 0 \end{aligned}$$

Remarque: On peut montrer que V_n n'est pas lui-même admissible. Déterminons 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.

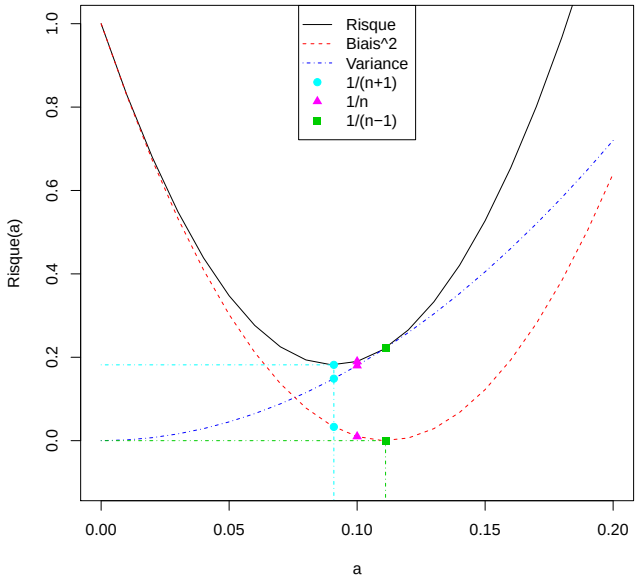
$$R(T_a) = \sigma^4[2a^2(n-1) + a^2(n-1)^2 + 1 - 2a(n-1)]$$

On dérive par rapport à a, d'où risque minimum pour $a = 1/(n+1)$. Le risque vaut alors $R(T_{1/(n+1)} = 2\sigma^4/(n+1)$ Cet estimateur est biaisé, et plus que l'estimateur empirique, mais de variance moindre ce qui compense. Il est uniformément meilleur que l'estimateur empirique qui n'est donc pas non plus admissible.

$$\frac{R(V_n) - R(T_{1/(n+1)})}{\sigma^4} = \frac{2n-1}{n^2} - \frac{2}{n+1} = \frac{n-1}{(n+1)n^2} > 0$$

Mais on peut préférer perdre un peu en risque pour gagner un estimateur sans biais, surtout quand la différence de risque est faible

$$\frac{R(\widehat{\sigma}^2) - R(T_{1/(n+1)})}{\sigma^4} = \frac{2}{n-1} - \frac{2}{n+1} = \frac{4}{n^2+1} > 0$$



5. *K_n est la somme de n v.a. indépendantes de variance 2. On peut écrire le TCL*

$$\sqrt{n-1}\left(\frac{K_{n-1}}{n-1}-1\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\mathrm{Var}(K_1)=2)$$

soit $\sqrt{n-1}(\widehat{\sigma}^2-\sigma^2)\xrightarrow{\mathcal{L}}\mathcal{N}(0,2\sigma^4)$ *ou encore* $\sqrt{n}(\widehat{\sigma}^2-\sigma^2)\xrightarrow{\mathcal{L}}\mathcal{N}(0,2\sigma^4)$

Si on ne fait plus l'hypothèse gaussienne, on peut écrire

$$\begin{aligned}\sqrt{n}\left(\frac{\sum_{i=1}^n(X_i-\bar{X})^2}{n}-\sigma^2\right)&=\sqrt{n}\left(\frac{\sum_i(X_i-\mu)^2}{n}-(\mu-\bar{X})^2-\sigma^2\right)\\&=\sqrt{n}\left(\frac{\sum_i(X_i-\mu)^2}{n}-\sigma^2\right)-\frac{1}{\sqrt{n}}\left(\sqrt{n}(\bar{X}-\mu)\right)^2\\&=\xrightarrow{\mathcal{L}}\mathcal{N}(0,\mathrm{Var}((X_1-\mu)^2))\end{aligned}$$

Le premier terme (TCL) tend vers $\mathcal{N}(0,\mathrm{Var}((X_1-\mu)^2))$. *Par le TCL,* $\sqrt{n}(\bar{X}-\mu)$ *tend* $\mathcal{N}(0,\sigma^2)$, *son carré (lemme de l'application continue) tend vers* $\sigma^2\chi^2(1)$, *et donc le deuxième terme tend vers* 0 *en proba. Le théorème de Slutsky permet de conclure. A distance finie, pour* n *grand, on a la loi approchée* $\widehat{\sigma}^2\overset{appr}{\sim}\mathcal{N}(\sigma^2,\mathrm{Var}((X_1-\mu)^2)/n)$

Exercice 2.
Un fabricant annonce que la contenance de ses boîtes de conserves est de $1kg$ avec un écart type de $10g$. Un organisme de consommateurs réalise avec les mêmes moyens de mesure une étude sur 16 boîtes de conserve prises au hasard et trouve une moyenne de $1,015kg$ et un écart-type corrigé de $15g$.

1. En supposant le modèle gaussien, rappeler la loi de l'estimateur $\bar{X}$ de l'espérance μ et celle de l'estimateur non biaisé de la variance $\widehat{\sigma}^2$.

2. Calculer la statistique de Student réalisée sur cet échantillon.

3. Quelle est la probabilité qu'un échantillon suivant les spécifications du fabricant ait une statistique de Student supérieure à celle observée sur l'échantillon tiré par l'organisme? Commenter.

4. Pour quel seuil s observe-t-on une probabilité de 5% de rencontrer une masse moyenne de l'échantillon de valeur inférieure à s?

5. Déterminer deux variables aléatoires $\widehat{\mu}_{min}$ et $\widehat{\mu}_{max}$ telles que

$$\frac{\widehat{\mu}_{min}+\widehat{\mu}_{max}}{2}=\widehat{\mu}\text{ et }\mathbb{P}(\widehat{\mu}_{min}\leq\mu\leq\widehat{\mu}_{max})=0.95$$

6. Une deuxième étude est faite avec la pesée de 40 boîtes prises au hasard. Le chargé d'études ne souhaite plus faire l'hypothèse gaussienne. Est-ce possible? Quelles réponses fera-t-il aux questions précédentes?

Correction. 1. *On suppose l'échantillon iid de loi marginale gaussienne* $\mathcal{N}(\mu, \sigma^2)$, *d'espérance* $\mu = 1$ *et de variance* $\sigma^2 = 0.01^2$. *La loi de* $\bar{X}$ *est gaussienne* $\mathcal{N}(\mu, \sigma^2/n)$. *L'estimateur non biaisé de la variance* $\widehat{\sigma}^2 = \sum_i (X_i - \bar{X})^2/(n-1)$ *suit la loi* $\sigma^2 \chi^2_{n-1}/(n-1)$ *où* χ^2_{n-1} *est la loi du Khi-deux à* $n-1$ *degrés de liberté.*

2. *La statistique de Student est définie par* $T_n = \sqrt{n}\frac{\bar{X}-\mu}{\widehat{\sigma}} \sim \mathcal{T}_{16-1}$. *Sa valeur sur l'échantillon est* $t_{obs} = \sqrt{16} \times (1.015 - 1)/0.015 = 4$

3. $\mathbb{P}(T_n > 4) = 1 - F_{\mathcal{T}_{15}}(4) = 0.0006$. *Si la masse des boîtes suit effectivement la loi annoncée par ce fabricant, et si on peut refaire* 10000 *fois l'expérience (tirage de* 10000 *échantillons de taille* 16, *calcul de la statistique de Student sur chacun), il y aura environ (et souvent pas exactement) six statistiques de Student observées ayant une valeur supérieure à* 4. *C'est donc un cas assez rare (mais pas improbable).*

4. *Si on considère* σ *inconnu, on cherche* s *tel que*

$$\mathbb{P}(\bar{X} \leq s) = 5\% = \mathbb{P}\left(\sqrt{n}\frac{\bar{X}-\mu}{\widehat{\sigma}} \leq \sqrt{n}\frac{s-\mu}{\widehat{\sigma}}\right)$$

Si q *est le quantile d'ordre* 5% *d'une loi de* $\mathcal{T}_{15}$, $q = -1.75$, *alors*

$$q = \sqrt{n}\frac{s-\mu}{\widehat{\sigma}}, \text{ soit } s = \mu + q\frac{\widehat{\sigma}}{\sqrt{n}}$$

La probabilité que le poids moyen soit inférieur à $s = 0.993$ *est de* 5%. *Si on utilise le fait que* σ *est connue,* $s^* = \mu + q^*\frac{\sigma}{\sqrt{n}} = 0.996$ *avec* $q^* = -1.64$ *le quantile de la gaussienne d'ordre* 5%.

5. *La loi gaussienne étant symétrique, on a*

$$\mathbb{P}(-q^* \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq q^*) = 0.95$$

où $q^* = 1.96$ *est le quantile d'ordre* 0.975 *de la loi gaussienne centrée réduite, d'où*

$$\mathbb{P}(\bar{X} - q^*\sigma/\sqrt{n} \leq \mu \leq \bar{X} + q^*\sigma/\sqrt{n}) = 0.95$$

L'intervalle observé est $[0.995, 1.005]$. *Si on considère la variance inconnue,*

$$\mathbb{P}(\bar{X} - q\widehat{\sigma}/\sqrt{n} \leq \mu \leq \bar{X} + q\widehat{\sigma}/\sqrt{n}) = 0.95$$

avec $q = 2.13$ *le quantile à* 0.975 *d'une loi de Student* $\mathcal{T}_{15}$ *et l'intervalle observé est* $[0.992; 1.008]$, *un peu plus large que le précédent, puisque l'incertitude est plus grande si* σ^2 *est inconnue.*

6. *Sans l'hypothèse gaussienne, on utilise le TLC pour définir une loi approchée ($n \geq 30$), soit si σ^2 est connue*

$$\mathbb{P}\left(\bar{X} - q^* \sigma/\sqrt{n} \leq \mu \leq \bar{X} + q^* \sigma/\sqrt{n}\right) \simeq 0.95$$

et si on la considère comme inconnue

$$\mathbb{P}\left(\bar{X} - q^* \widehat{\sigma}/\sqrt{n} \leq \mu \leq \bar{X} + q^* \widehat{\sigma}/\sqrt{n}\right) \simeq 0.95$$