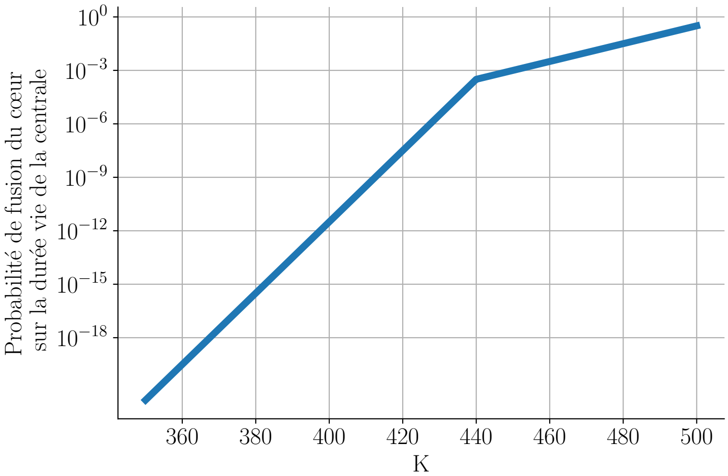
TD 3 : Phénomènes statistiques et applications

Conception d'une centrale nucléaire

Dans cette exercice, on considère une situation et des données fictives afin d'illustrer des situations communément rencontrées en science des données.

Les concepteurs d'une centrale nucléaire veulent éviter le scénario catastrophe d'une fusion du cœur de la centrale. La probabilité de fusion du cœur est déterminée par la capacité K du système de refroidissement, selon la loi décrite ci-dessous.



Les concepteurs ont le choix entre les technologies vendues par les sociétés *Nuke and cranny* et *Fusion impossible*. La capacité de refroidissement pour ces technologies n'est pas connue théoriquement et doit être estimée empiriquement à partir de mesures très bruitées, mais sans biais. Par sans biais, on veut dire que pour chaque système on peut modéliser les mesures comme des tirages aléatoires indépendants d'une distribution de probabilité dont l'espérance est la capacité de refroidissement réelle du système.

Supposons que l'on obtenu 8 mesures pour le système de la société *Fusion impossible* :

$$X_F = 11.5\ 0.02\ 779\ 13.9\ 1124\ 526\ 0.0005\ 594$$

et 7 mesures pour le système de la société *Nuke and cranny* :

$$X_N = 488\ 631\ 110\ 432\ 444\ 705\ 178.$$

Comme vous pouvez le vérifier par vous-même, la moyenne de l'échantillon X_F est environ $\hat{K}_F = 381$ et celle de l'échantillon X_N est environ $\hat{K}_N = 427$.

On notera P_F et P_N les distributions de probabilité de laquelle ces échantillon sont respectivement tirés (notez que la seule chose que l'on sait sur ces distribution est qu'elles ont produits les échantillons obtenus) et K_F et K_N les capacité réelles des systèmes de refroidissement.

Jo Lerigolo propose d'estimer les risques encourus en prenant la moyenne de chaque échantillon et en utilisant le graphe ci-dessus pour estimer le niveau de risque correspondant. Si les deux systèmes permettent d'atteindre un niveau de risque raisonnable, il propose de sélectionner celui dont le niveau de risque est le plus bas d'après sa méthode.

1. Implémentez la solution proposée par Jo et indiquez pour chaque société si la technologie proposée semble suffisamment performante. Quelle société devrait-on préférer selon Jo ?

2. Que pensez-vous de l'approche proposée par Jo ?

3. Pour mieux évaluer les risques encourus, il faudrait que nous ayions une idée d'à quel point les capacités estimées $\hat{K}_F$ et $\hat{K}_N$—auxquelles nous avons accès—peuvent être éloignée des capacités réelles K_F et K_N—qui sont celles qui nous intéressent réellement. Supposons provisoirement que nous connaissions la distribution P_F. Pour faire simple, disons qu'une valeur tirée selon P_F vaut 379 avec probabilité 1/2 et 383 avec probabilité 1/2. Proposez une méthode pour évaluer à quel point la capacité estimée $\hat{K}_F$ peut être éloignée de la capacités réelle K_F.

4. Par groupe de deux ou seul, tirez aléatoirement 4 échantillons de taille 8 de P_F, calculez la moyenne de chaque échantillon et communiquez vos résultats à l'enseignant.

5. A partir des résultats de tous les groupes, représentez graphiquement l'estimation obtenue pour la distribution de $\hat{K}_F$. D'après ce graphe, doit-on s'attendre à ce que $\hat{K}_F$ soit très éloigné de K_F ?

6. La réponse donnée à la question précédente permet-elle de répondre à la question initiale ?

7. En pratique nous n'avons accès à P_F que par le biais de l'échantillon X_F. Proposez et représentez graphiquement une estimation $\hat{P}_F$ de P_F obtenue sur la base de X_F. Justifiez votre réponse sur la base d'un résultat vu en cours.

8. Comment pouvez-vous tirer aléatoirement un échantillon de votre estimation $\hat{P}_F$ de P_F en pratique ?

9. Reprenez les questions 4 et 5 en utilisant votre estimation $\hat{P}_F$ comme un substitut à P_F. Cette idée de remplacer la distribution théorique par une estimation empirique pour analyser la distribution d'un estimateur est célèbre en statistique. Elle porte le nom de *bootstrapping* et a été introduite et étudiée par Bradley Efron en 1979.

10. Les tailles d'échantillons considérées ici sont pratiques pour se familiariser avec les concepts clés en faisant les calculs et graphes à la main, mais elles sont trop petites pour que $\hat{P}_F$ constitue une bonne estimation de P_F et il faut donc être très prudent quand à l'interprétation des résultats. Pour pouvoir répondre de manière fiable à la question posée par la méthode du bootstrap, les concepteurs de la centrale nucléaire ont besoin de plus de données (il faut aussi qu'ils tirent plus d'échantillons de $\hat{P}_F$ et $\hat{P}_N$ que nombre-de-groupe $* 4$ pour éviter que le résultat ne dépende trop du côté aléatoire de ces tirages). Ci-dessous, on donne les résultats obtenus en reproduisant les étapes des questions précédentes pour des échantillons de P_F et P_N de taille 5000 au lieu de 8 et en réalisant 1000 tirages aléatoires d'échantillons de taille 5000 de $\hat{P}_F$ et $\hat{P}_N$ au lieu de nombre-de-groupe $* 4$ tirages aléatoires d'échantillons de taille 8. Sur la base de ces résultats, y a-t-il une technologie satisfaisante pour la construction de la centrale ? Comparez avec les résultats obtenus par la méthode de Jo Lerigolo (pour les échantillons considérés on a $\hat{K}_F \approx 409$ et $\hat{K}_N \approx 415$).

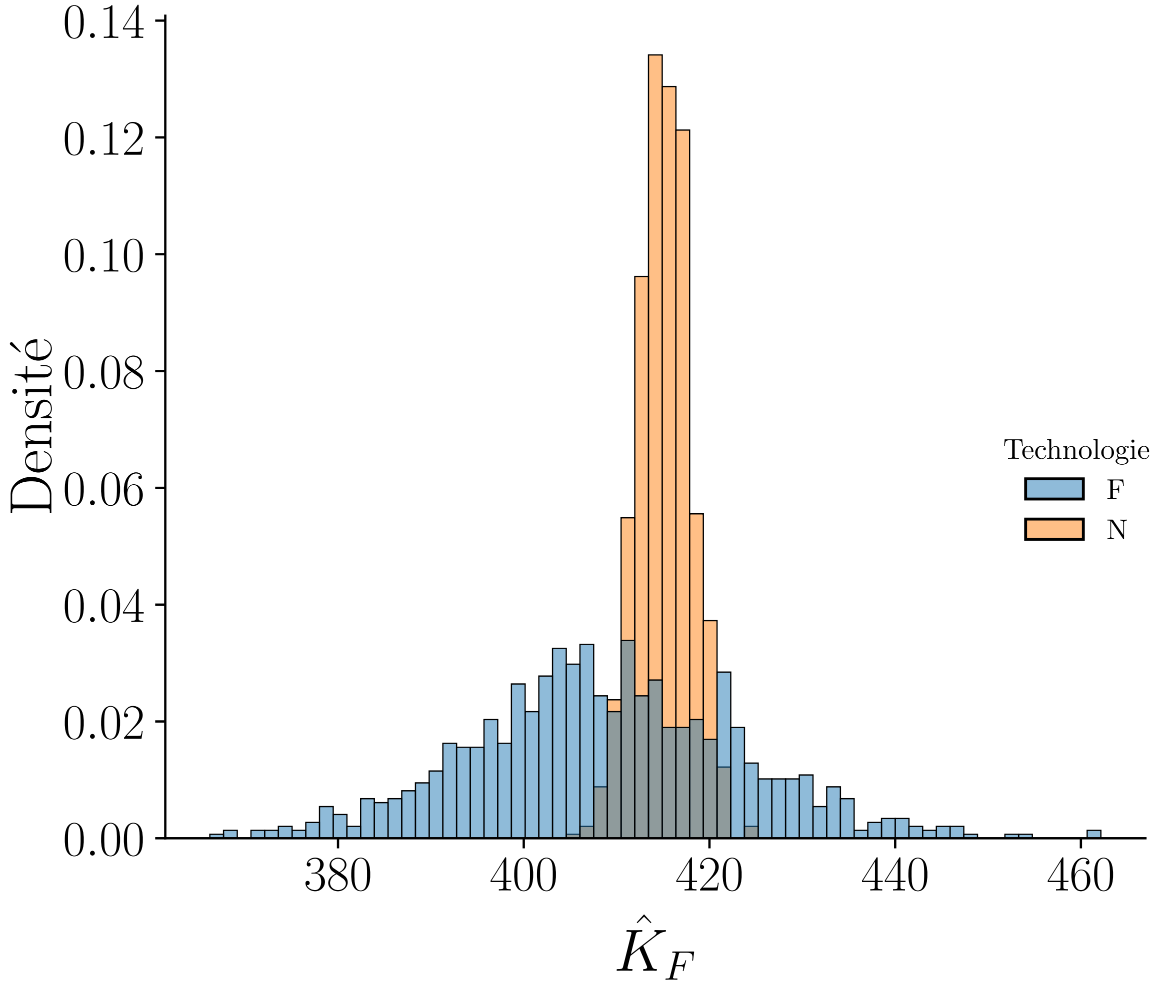


Figure 2 – Estimation de la distribution empirique de $\hat{K}_F$ obtenue par bootstrapping.