TP4 - Statistique avec R - Probabilités et inférence statistique

---

OBJECTIFS DU TP :

Le but de ce TP est de se familiariser aux lois de probabilités sous R ainsi qu'à plusieurs concepts que vous reverrez en statistique inférentielle.

Travaillez soit sur le cluster R soit sur le Datalab/Onyxia. Vous pouvez aussi librement choisir de coder sur un document Quarto `.qmd` (au sein de *chunks* donc) ou sur un scrpit `.R` classique.

Commencer par charger votre Rprojet `Stats_avec_R` créé au TP1, puis affecter un nouveau dossier "TP3 fonctions". Pour importer vos données, cliquez sur le bouton `Upload` situé dans l'onglet `Files`, en dessous de votre environnement.

Exercice 1

Première partie : la loi des grands nombres

On ici souhaite illustrer la loi des grands nombres.

> **Rappel : Loi des grands nombres**
>
> Soient $X_1, X_2, \dots, X_n$ une suite n de variables aléatoires indépendantes et identiquement distribuées (i.i.d) ie. un échantillon de n observations indépendantes de X. Alors, la moyenne empirique (observée) $\overline{X_n}$ converge asymptotiquement vers la moyenne théorique $\mu := \mathbb{E}[X]$, lorsque la taille n de l'échantillon augmente.

1. Rappeler quelle est l'espérance d'une loi exponentielle de paramètre λ.

2. Générer 3 échantillons de tailles différentes constitués de variables aléatoires suivant une loi exponentielle de paramètre $\lambda = 1$:

— un échantillon de taille $n = 10$
— un échantillon de taille $n = 30$
— un échantillon de taille $n = 1000$

Vous précéderez votre code de `set.seed(1234)` afin de fixer la graine aléatoire.

3. Calculer la moyenne sur chacun des échantillons. Comparer les moyennes obtenues à l'espérance théorique, que constatez-vous ?

4. Représenter les histogrammes de chacun des échantillons. L'axe des ordonnées correspondra à la densité de probabilité (et non pas aux effectifs).

5. Calculez la médiane théorique d'une loi exponentielle de paramètre $\lambda = 1$ (indication : pensez à la fonction pour calculer les quantiles) ? Calculez ensuite les medianes des 3 échantillons précédents. Que remarquez-vous ?

Deuxième partie : le théorème central limite

On souhaite maintenant illustrer le théorème central limite (TCL).

Pour rappel : selon le TCL, la somme de n variables aléatoires i.i.d (de carrés intégrables) converge vers une loi normale lorsque n est grand.

> **Rappel : le Théorème Central Limite (TCL)**
>
> Soient $X_1, X_2, \dots, X_n$ une suite n de variables aléatoires indépendantes et identiquement distribuées (i.i.d) ie. un échantillon de n observations indépendantes de X, de carrés intégrables, d'espérance μ et de variance σ^2. Alors, lorsque la taille n de l'échantillon augmente :
>
> $$\sqrt{n}\,\frac{\overline{X_n}-\mu}{\sigma} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} Z \quad \text{où} \quad Z \backsim \mathcal{N}(0,1)$$

6. En question 2 vous avez créé un échantillon constitué de 1 000 variables aléatoires. Maintenant, créer non pas 1 (comme en question 2) mais 1 000 échantillons de 1 000 variables aléatoires suivant une loi exponentielle de paramètre $\lambda = 1$. Pour chaque échantillon, vous calculerez la somme totale des réalisations (observations). Stockez le calcul des totaux dans un vecteur pour obtenir un vecteur de 1 000 totaux.

7. Représenter l'histogramme associé à votre vecteur de sommes. À quelle loi vous fait penser la forme de la distribution ?

8. À partir de votre vecteur de sommes, créer un vecteur de moyennes. Représenter ensuite la distribution de votre vecteur de moyennes. Quelle distribution est attendue ? Quelle en seraient les paramètres ?

9. Superposer à votre histogramme précédent, la fonction de densité de la loi qui vous semble correspondre.

10. Quelles sont les quantiles 0,025 et 0,975 associés à une loi $N(1, (1/\sqrt{1000})^2)$? Dit autrement, entre quelles valeurs se trouvent 95 % des observations pour une telle loi ?

11. Retrouver les quantiles précédents en partant d'une loi normale centrée réduite. On travaille donc cette fois sur notre vecteur des moyennes, noté $\overline{X_n}$, que l'on centre et réduit, soit

    $$Z = \sqrt{1000}\,\frac{\overline{X_n}-1}{1}$$

    avec $Z \sim N(0,1)$. On sait donc que $P(-1,96 < Z < 1,96) \approx 0.95$. Remarque : cette question pourrait être faite à la main, sans utiliser `R`.

Troisième partie : introduction aux intervalles de confiance

Cette partie est une partie d'application : on utilise cette fois un jeu de données.

12. Charger la librairie `MASS` et les données *quine* en faisant :

```
library(MASS)
data(quine)
```

Aller dans le `help.start()` puis taper "*quine*" pour découvrir un descriptif des données. On travaille avec les données concernant le nombre de jours d'absence.

13. Représenter l'histogramme de la variable `Days`. A quelle distribution, cela vous fait penser ?

14. Essayer de représenter la distribution théorique qui vous semble correspondre à cet histogramme. À vous d'en trouver le paramètre ! Sur quoi pouvez-vous vous appuyer pour le trouver ?

15. Quel est le quantile 0.95 d'une telle loi théorique ? On considère qu'un élève doit être exclu si le nombre d'absences dépasse ce quantile. Combien d'élèves sont concernés dans notre jeu de donnés ?

16. Au vu de ce qui a été fait en 2e partie, on sait que

$$\sqrt{n}\ \frac{\overline{X_n}-\mu}{\sigma} \sim N(0,1)$$

. Ici on ne connait ni μ ni σ. On pose alors

$$T = \frac{\overline{X_n}-\mu}{S}\sqrt{n} \sim St(n-1)$$

avec

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2}$$

l'estimateur empirique de σ. Quels sont les quantiles 0.025 et 0.975 d'une telle loi de Student ?

17. On fait ici une introduction aux intervalles de confiance que vous verrez en cours de statistique inférentielle. On suppose que tous les élèves n'ont pas été échantillonnés. On souhaite connaître le nombre moyen de jours d'absence au sein de l'établissement.

À partir de la question précédente et en vous inspirant de ce qui a été fait en question 10, trouver les valeurs x_1 et x_2 telles que $P(x_1 < \mu < x_2) \approx 0.95$. Dit autrement, dans quel intervalle peut-on encadrer μ avec une probabilité de 0.95 ? En déduire un intervalle pour le nombre moyen de jours d'absence au sein de cet établissement.

Quatrième partie : influence de la taille d'échantillon

On continue à travailler sur la variable `Days` de la table quine avec

$$T = \sqrt{n}\ \frac{\overline{X_n}-\mu}{S} \sim St(n-1)$$

18. En s'inspirant de la question précédente, créer une fonction qui calcule l'intervalle de confiance à 95% pour la moyenne. La fonction retournera un vecteur de 2 valeurs : une borne inférieure et une borne supérieure.

Cette fonction prendra uniquement en paramètres :

— un vecteur numérique (les données)
— la taille de l'échantillon (n).

Le calcul de l'intervalle sera fait en fonction de ce paramètre n. Par exemple, si n=10, le calcul sera fait en utilisant uniquement les 10 premières observations de la table (sans tri), comme si la table ne contenait que 10 observations. Si $n = 100$, on ne prend que les 100 premières…etc. Le paramètre n ne peut évidemment pas être supérieur au nombre total d'observations (ici 146). Vous rajouterez donc un test bloquant (`stop`) pour que n ne dépasse pas le nombre total d'observations.

Tester votre fonction en créant :

— int10 : l'intervalle pour $n = 10$
— int50 : l'intervalle pour $n = 50$
— int : l'intervalle avec toutes les observations

19. Représenter l'ensemble des données de la variable `Days` avec la fonction `plot()`. Superposer les bandes de confiance de `int10` (couleur rouge), de `int50` (couleur bleu) et de `int` (couleur vert). Vous rajouterez également une droite en noir d'équation `y=mean(quine$Days)`. Indication : pour superposer les droites, vous pouvez utiliser la fonction `abline()`. Que constatez-vous ?