TP 4: Loi de grandes nombres. Théorème central limite.

anna.melnykova@univ-avignon.fr

La séance d'aujourd'hui est consacrée aux propriétés statistiques d'un échantillon de grande taille. Notamment, on s'interesse à l'application de la loi de grandes nombres et le théorème central limite. Pour rappel, grace à la loi de grandes nombres, on a la garantie théorique que la moyenne empirique d'une sequence de variables aléatoires se converge vers l'espèrance théorique. En effet, considerons une suite de variables i.i.d. $X_1, \dots, X_n$, avec l'espèrance μ et la variance σ^2. La moyenne empirique $\bar{X}_n$ de cette suite de variables possède les propriétés suivantes:

$$\mathbb{E}[\bar{X}_n] = \mu \qquad Var[\bar{X}_n] = \frac{\sigma^2}{n}$$

Grace à l'inégalité de Tchebychev on a:

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon}$$

Autrement dit, quand $n \to \infty$, la difference entre la moyenne empirique et l'espèrance devient negligeable quand la taille d'échantillon augmente. En plus, grace au Théorème Centrale Limite, la distribution de la moyenne empirique peut être approché par la loi Normale pour n suffisament grand:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Petite astuce: sur la page du cours, vous avez le script R qui sert à illustrer le TCL. Vous pouvez l'adapter pour traiter les exercices suivantes.

Loi uniforme discrète

Dans cet expérience, on lance plusiers dés à 6 faces. On cherche à savoir quelle est la répartition d'une somme de dés.

1. D'abord, considérons 2 dés. Quelles sont les modalités de variable, qui décrit la somme de faces? Est-ce que toutes les modalités sont équiprobables? Quelle est l'espérance de cette variable? Quelle est l'espérance de lancer de 10 dés? 100?

2. Quelle est l'espérance et la variance de la moyenne de 2 dés? 10 dés? 100?

3. Faisons une petite experience. Rappelez-vous du TP 1 (Exercice 3). On peut simuler $n = 10$ lancés d'un dé avec un programme suivant:

```
n <- 10
X <- floor(6*runif(n))+1
```

3. Simulez la moyenne de X. Est-ce que la valeur observée correspond à son espérance théorique?

4. Maintenant, on cherche à tracer la distribution de la variable 'moyenne de 10 dés'. Pour ça, simulons cette variable 100 fois et traçons la distribution avec une histogramme:

```
k <- 100       # nombre de simulations
S <- numeric(k) # vecteur pour stocker les sommes
for (i in 1:k){
  # Ici, ajoutez votre code pour simuler 100 dés et les stocker dans un vecteur X
  S[i] <- mean(X)
}
hist(S, prob = T)
```

5. Maintenant, il nous reste de vérifier le TCL. Rappeler-vous de la formulation de TCL. Comment doit-on transformer le vecteur S pour qu'il soit approché par la loi normale centrée réduite?

6. Faites la transformation et réalisez une histogramme. Qu'est-ce que vous observez? Est-ce qu'on peut plutôt dire que la distribution est normale? Augmentez la taille d'échantillon et répétez.

```
S_centre  # faites la transformation ici
hist(S_centre, prob = T)
curve(dnorm(x), col = "red", lwd = 2, add = T)
```

Loi continue

On va vérifier si le théorème centrale limite s'applique également à la moyenne empirique d'une loi exponentielle. On suppose que le temps (en heures) entre deux pannes d'un serveur suit une loi exponentielle de paramètre $\lambda = 1/5$, c'est-à-dire que la durée moyenne entre deux pannes est de 5 heures.

1. Générez $n = 1000$ valeurs suivant une loi exponentielle de paramètre $\lambda = 1/5$, en utilisant la fonction `rexp`. Faites un histogramme de ces valeurs et comparez avec la densité théorique (en superposant la courbe de la densité sur l'histogramme) pour vérifier qu'il s'agit bel et bien de la loi exponentielle.
2. En prenant $n = 100$, calculez $k = 1000$ moyennes empiriques de cette loi et représentez-les par un histogramme. Changez $n = 10$ et $n = 10000$. Quoi observez-vous?

3. Par quelle loi peut-on approcher la distribution de la moyenne empirique? Quels sont les paramètres de cette loi? (Question du cours)

4. En suivant l'exemple de l'exercice précédant, transformez le vecteur des moyennes empiriques pour obtenir une **loi normale centrée réduite**. Vérifiez bien que votre code marche pour différentes valeurs de n. Quoi passe-t-il si on augmente k?

Loi de Cauchy

Maintenant, on va considérer le cas ou la loi de grandes nombres (ni TCL) ne marche pas. Souvenez-vous de la loi de Cauchy, qui est défini comme suite. Soit X et Y deux variables indépendantes normales centrées réduites. Alors,

$$Z = \frac{X}{Y} \sim Cauchy(0, 1).$$

Pour simuler cette variable, on peut simuler 2 échantillons de la loi normale de la même taille, et puis calculer la relation.

```
n <- 100
X <- rnorm(n)
Y <- rnorm(n)
Z <- X/Y
# hist(Z)
summary(Z)
var(Z)
```

1. Relancez le code plusieurs fois. Qu'est-ce que vous observez? Est-ce qu'on obtient (un peu près) la même moyenne? Variance? Médiane?

2. Augmentez n et reesayez.

3. Maintenant, on va essayer de construire un éstimateur de la moyenne en utilisant plusieurs échantillons. Pour ça, on va simuler $k = 1000$ échantillons de taille $n = 10000$ chacun et puis sauvegarder les moyennes empiriques:

```
k = 10000
n = 10000
z_hat <- numeric()
for (i in 1:k){
  X <- rnorm(n)
  Y <- rnorm(n)
  Z <- X/Y
  z_hat[i] <- mean(Z)
}
var(z_hat)
summary(z_hat)
```

4. Relancez le code plusieurs fois. Est-ce que vous avez les mêmes quartiles? Et la moyenne/variance? Quelle conclusion peut-on tirer?

5. *Question bonus:* avec la fonction `quantile(z_hat,alpha)` (ou $\alpha \in (0, 1)$) vous pouvez explorer les quantiles empiriques de n'importe quelle ordre de votre estimateur. Essayez de visualiser la distribution des observations qui sont dans l'intervalle $[q_{0.05}, q_{0.95}]$ (avec une histogramme ou une boîte à moustaches, par exemple). À quoi ressemble cette distribution?