INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN
L. Noiret,
A. Rogozan

TP
Théorème Central Limite et applications

asi Architecture des Systèmes d'Information
3e année

Dans ce TP, nous allons mettre en pratique le théorème central limite (TCL) et regarder les caractéristiques de la distribution de la moyenne empirique. Le TCL est particulièrement important pour la mise en oeuvre de tests statistiques et la construction d'intervalles de confiance de la moyenne.
Il nous dit que la distribution de la distribution moyenne empirique de n'importe quelle suite de variables aléatoires iid (indépendantes et identiquement distributées) suit une loi gaussienne lorsque la taille de l'échantillon est suffisamment grande.

1 Plus l'échantillon est grand, moins la moyenne varie

La moyenne est un indicateur de tendance central d'une distribution. Elle est donc souvent utilisée dans les tests statistiques afin de comparer deux populations.

Soient $X_1..X_n$ une suite de variables aléatoires i.i.d. de moyenne μ et de variance σ^2. On considère la statistique suivante (estimateur empirique de la moyenne)

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

.

1. Donner l'espérance mathématique de $\bar{X}_n$ en fonction de μ.
2. Donner la variance de $\bar{X}_n$ en fonction de σ^2. En déduire l'écart type de $\bar{X}_n$.
3. En déduire l'espérance et la variance de $\bar{Z}_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$

Rappels :

- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
- $Var(aX + b) = a^2 Var(X)$
- $\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$
- $Var(X_1 + X_2) = Var(X_1) + Var(X_2) + Cov(X_1, X_2)$
- si X_1 et X_2 sont indépendantes, $Var(X_1 + X_2) = Var(X_1) + Var(X_2)$.

Application

Une entreprise de vêtement américaine *MadeInC* souhaite réaliser une enquête sur la taille des françaises afin d'ajuster la taille de ces vêtements aux clientes de l'hexagone. Pour cela elle recrute 50 enquêteurs dans toute la France afin de mesurer des femmes.

3. On sait (via une enquête secrètement réalisée par un concurrent) que la taille des clientes suit une loi normale de moyenne 162.5 cm et de variance 2. Simuler dans Matlab une population de 1000 clientes (fonction *normrnd*), et représenter leur distribution (estimateur à noyau de la densité: fonction *ksdensity*). Calculer la moyenne empirique $\bar{x}_n$ et l'écart type empirique s de votre population. Pourquoi ne retrouvez-vous pas exactement 162.5 et $\sqrt{2}$?

4. Chaque enquêteurs mesure $n = 30$ femmes. Simuler les résultats des 50 enquêteurs. Calculer la taille moyenne empirique ($\bar{x}_n$) obtenue par chaque enquêteur. Vous venez de créer un échantillon de 50 moyennes de tailles.

5. Ajouter la distribution des moyennes au graphique précédent (*hold on ... hold off*). Commenter les résultats.

6. Comment expliqueriez-vous à vos parents (avec des mots), le fait que la dispersion de la moyenne est plus petite que la dispersion de la population?

7. Un des enquêteurs de *MacInC* est un peu étourdi (ou mal organisé). Il a réalisé des enquêtes dans différents pays, mais il a oublié d'inscrire le pays associé à chaque enquête. Il trouve un fichier avec 50 mesures. La moyenne des données de ce fichier est 164 et de variance 2. Cette enquête a-t-elle été réalisée en France?

2 TCL

L'entreprise d'ampoules Neaugreen souhaite estimer la durée de vie de nouvelles ampoules grand luxe en cristal à 45 euros l'unité. Pour cela, elle demande à 40 de ses succursales de réaliser des tests d'usure. Dans chaque succursale, les employés doivent maintenir $n = 20$ ampoules allumées en permanence et enregistrer la durée de vie de chaque ampoule. A l'issue de l'enquête, le chargé détudes statistiques analyse les résultats. Cette enquête permis de montrer que la durée de vie des ampoules suivent une loi exponentielle de paramtre $\lambda = 5$ correspondant une durée de vie moyenne de $\frac{1}{\lambda} = 0,2$ année (2,4 mois) et une variance variance $\frac{1}{\lambda^2}$ de 0,04 années.

1. Simuler un échantillon de 1200 ampules suivant une loi $\mathcal{E}(5)$ (fonction *exprnd*, attention le paramètre attendu dans Matlab est la moyenne et non λ) et afficher l'histogramme.

2. Simuler les résultats des 40 succursales (simulation n = 20 ampoules/succursale). Calculer la durée de vie moyenne de chaque succursale $(\bar{x}_{30}^1, ..., \bar{x}_{30}^{40})$.

3. Afficher l'histogramme de l'échantillon des moyennes. Calculer la moyenne et l'écart-type de la moyenne. Commenter.

4. Le chargé détude statistiques est un peu étonné par la forme de l'histogramme obtenu, et demande aux 40 succursales de refaire les tests, mais cette fois avec 100 ampoules. Répétez l'expérience et visualiser les résultats l'aide d'un histogramme.

5. Interprêter les résultats de simulation à l'aide du théorème central limite.

6. question BONUS : Une ampoule LED à une durée de vie moyenne de 6 ans (distribution exponentielle). A votre avis, la durée de vie moyenne des ampoules Neaugreen est-elle différente des ampoules LED?

   1. Quelle est distribution de $\frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}}$?

      Rappels :

      - estimateur de la variance : $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$
      - $(n-1)\frac{S^2}{\sigma^2}$ suit une loi du Chi deux à $n-1$ degré de liberté $\mathcal{X}^2(n-1)$
      - Soit Z une variable gaussienne $\mathcal{N}(0,1)$ et U une variable $\mathcal{X}^2(k)$, Z et U indépendantes. Alors $\frac{Z}{\sqrt{\frac{U}{k}}}$ suit une loi de student à nk degrés de liberté.

   2. Quelle est la probabilite qu'une ampoule Neaugreen ait une durée de vie moyenne :

      - inférieure à 1 mois?
      - superieure à 12 mois

   3. Donner un intervalle où se trouve 95% des durée de vie moyennes? Y-a-t-il une seule solution?

   4. Conclure.