TP 5: Intervalles de confiance. Tests statistiques.

anna.melnykova@univ-avignon.fr

Exercice 1

En 2017, la population active en France a été estimée à 29.7 millions personnes. Dans ce nombre, on compte aussi les gens au chomage, soit 2.9 millions. On va simuler la population totale en France à 2017 et faire une 'étude' de taux de chomage.

```
Pop17 <- rep(0,29700000) # Population active
Pop17[1:2900000] <- 1 # On remplace les 2.9 millions d'élèments par 1 pour designer les chomeurs
```

1. Quelle loi suit la variable 'nombre de personnes à chomage' dans la sous-population de taille k? Avec quel(s) paramètre(s)?

2. Calculez la moyenne du vecteur `Pop17`. À quoi correspond cette moyenne? Sauvegardez-le dans la variable `taux`.

3. On se place dans le role d'un institut qui fait un sondage dans la population française pour determiner le taux de chomage. Pour ça, on interroge 100 personnes et sauvegarde les résultats dans un vecteur:

```
n = 100
# commande "sample" fait le tirage de n élèments du vecteur Pop17
Sondage17 <- sample(Pop17, n, replace = FALSE)
```

4. Calculez la moyenne du vecteur `Sondage17`. Est-ce que la moyenne est égale à `taux`?

Maintenant, on va construire une intervalle de confiance de 80%. Souvenez-vous que pour la proportion, l'intervalle de confiance de $1-\alpha\%$ est donné par la formule suivante:

$$\left[\hat{p}_n - q_{1-\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + q_{1-\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right],$$

ou $q_{1-\alpha/2}$ c'est la quantile de la loi normale centrée réduite. Les quantiles de la loi normale centrée réduite on calcule avec la fonction `qnorm`.

5. Pour implementer l'IC dans R, calculez la borne infèrieure et superieure en se basant sur le taux de chomage éstimé par le sondage:

```
alpha <- 0.2
ICInf <-
ICSup <-
IC <- c(ICInf, ICSup)
IC
```

6. Dans R, on peut aussi calculer cet intervalle de façon exacte, en utilisant la loi binomiale (souvenez-vous que la formule pour IC se base sur le théorème centrale limite) avec la commande suivante. Est-ce que le résultat obtenu correspond à l'IC obtenue avec l'approximation par la loi normale?

```r
prop.test(sum(Sondage17),n, conf.level = 0.8)$conf.int
```

7. Est-ce que le vrai taux de chomage se trouve dans l'IC obtenue? Essayez de relancer le code plusieurs fois en utilisant l'autre échantillon (i.e. relancez les commandes à partir de `sample`) et commentez le résultat.

8. Augmentez la taille d'échantillon et commentez. Est-ce que la probabilité que l'IC contient le vrai taux de chomage a changé? Qu'est-ce qui est changé?

9. Finalement, on va construire 20 intervalles de confiance et les visualiser sur la même graphique. Commentez le résultat. Est-ce que toutes les intervalles contiennent la vraie valeur du taux de chomage? Pourquoi?

```r
k <- 20
ConfInts <- matrix(ncol = k, nrow = 2)
for (i in 1:k){
 Sondage17 <- sample(Pop17, n, replace = FALSE)
 ConfInts[,i] <- prop.test(sum(Sondage17),n, conf.level = 0.8)$conf.int[1:2]
}
matplot(ConfInts,rbind(1:k,1:k),type="l",lty=1, xlab = "Intervalles de confiance", ylab = "")
abline(v = mean(Pop17), lwd = 2, col = "red")
```

10. Repetez l'expérience (à partir de la question 3) en augmentant le nombre de personnes interrogées (par exemple, $n = 1000$) et commentez.

Exercice 2

En 2021 le nombre de gens inscrites à Pole Emploi s'établit à 5.37 millions, tandis que la population active compte 28.9 millions personnes.

1. Simulez la population active et les chômeurs en utilisant l'exemple de l'Exercice 1. Stockez-la dans la variable `Pop21`.
2. Prenez l'échantillon de 1000 personnes dans la population totale et proposez l'intervalle de confiance de 90% pour déterminer le taux de chomage à 2021. Comparez-la avec l'intervalle de confiance du même seuil pour le taux de chomage à 2017.

3. Finalement, on va faire le test statistique en prenant la marge d'erreur 10% sur l'échantillon `Sondage21` pour déterminer si le taux de chomage est différent de celui à 2017 (9.8%, i.e., $p = 0.098$). Formulez matematiquement les hypothèses de test. Pour exécutez, on utilise les commandes suivantes (variable `taux` est celui déclarée dans la question 2):

```r
prop.test(sum(Sondage21),n, p = taux, conf.level = 0.9)
```

4. Quelle est la conclusion du test? Est-ce que le taux de chomage est différent de celui à 2017?

Exercice 3

Un administrateur système souhaite analyser le temps entre deux pannes d'un serveur. Il suppose que ces temps suivent une loi exponentielle de paramètre λ, c'est-à-dire que le temps moyen entre deux pannes est donné par $1/\lambda$.

Il collecte un échantillon de $n = 50$ durées (en heures) entre pannes et veut :

- Estimer l'intensité λ à partir des données en utilisant la méthode des moments.
- Construire un **intervalle de confiance asymptotique** pour λ.
- Tester l'hypothèse $H_0 : \lambda = 1/5$ contre $H_1 : \lambda > 1/5$ au seuil de 5%.

On suppose que les durées entre pannes suivent une loi exponentielle et que l'on dispose de l'échantillon suivant :

```r
set.seed(1984)
n <- 50
lambda_vrai <- 1/5  # Valeur théorique de lambda
data <- rexp(n, rate = lambda_vrai)  # Génération des données
```

1. Souvenez-vous comment on peut estimer le paramètre λ à partir des données. Sauvegarder l'estimation dans une variable `lambda_hat`.

Maintenant, on va construire un intervalle de confiance pour le paramètre λ. Pour rappel, pour un grand n, l'estimateur $\hat{\lambda}$ (que vous avez défini dans la première question, normalement) suit une **distribution normale asymptotique** :

$$\hat{\lambda} \approx \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$$

Un **intervalle de confiance asymptotique à** $(1-\alpha)\%$ pour λ est alors donné par :

$$\left[\hat{\lambda} - q_{1-\alpha/2} \times \frac{\hat{\lambda}}{\sqrt{n}}, \quad \hat{\lambda} + q_{1-\alpha/2} \times \frac{\hat{\lambda}}{\sqrt{n}}\right]$$

2. Construire un intervalle de confiance de 95% comme dans l'Exercice 1, en utilisant les données simulées (vecteur `data`).

Maintenant, on va construire un test d'hypothèse, en utilisant la variable suivante:

$$Z = \frac{\hat{\lambda} - \lambda_0}{\hat{\lambda}/\sqrt{n}}$$

3. Quelle est la loi, suivie par cette variable, si l'hypothèse H_0 est correcte?

4. Calculer la valeur de cette variable sur vos données et stockez-le dans une variable `z_obs`. En utilisant la fonction `pnorm`, calculer la probabilité d'observer une valeur qui dépasse `z_obs`, si l'hypothèse H_0 est correcte. Est-ce que cette probabilité est plutôt grande, ou plutôt petite? Quelle conclusion peut-on faire?

5. Reprenez les questions (1)-(4) pour des données simulées avec le code suivant:

```
set.seed(1984)
n <- 50
lambda2 <- 2/5  # Valeur théorique de lambda
data2 <- rexp(n, rate = lambda2)  # Génération des données
```

6. Est-ce que l'intervalle de confiance contient la valeur 1/5? À quelle conclusion arrive-t-on dans le test?