TP 4: 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 chômage.

```
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 chômage' dans la sous-population de taille k? Avec quel(s) paramètre(s)?

Soit X - nombre de personnes au chomage. X suit une loi binomiale de paramètres $n = 29.7$ millions et $p = 29700000/2900000 \approx 0.1$. Alors, $X \sim Bin(29700000, 0.1)$

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

```
taux = mean(Pop17)
# Probabilité qu'un individu prit au hasard soit au chomage
print(mean(taux))
```

```
## [1] 0.0976431
```

```
# Il s'agit du vrai taux car on "connait" toute la population
```

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

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

```
tauxSondage = mean(Sondage17) # Moyenne du vecteur Sondage17
print(mean(tauxSondage))
```

```
## [1] 0.09
```

Avec $n = 100$, la moyenne du vecteur Sondage17 est plutôt éloignée du taux. Cela s'explique car on n'a pas interrogé toute la population.

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 le quantile de la loi normale centrée réduite. Les quantiles de la loi normale centrée réduite on calcule avec la fonction `qnorm`.

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

```
alpha <- 0.2
ICInf <- tauxSondage - qnorm(1-alpha/2)*sqrt(tauxSondage*(1-tauxSondage)/n)
ICSup <- tauxSondage + qnorm(1-alpha/2)*sqrt(tauxSondage*(1-tauxSondage)/n)
IC <- c(ICInf, ICSup)# vecteur intervalle de confiance
print(IC)
```

```
## [1] 0.05332433 0.12667567
```

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?

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

```
## [1] 0.05562687 0.13938645
## attr(,"conf.level")
## [1] 0.8
```

Pratiquement toutes les valeurs obtenues dans la question 3 sont compris dans l'intervalle.

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.

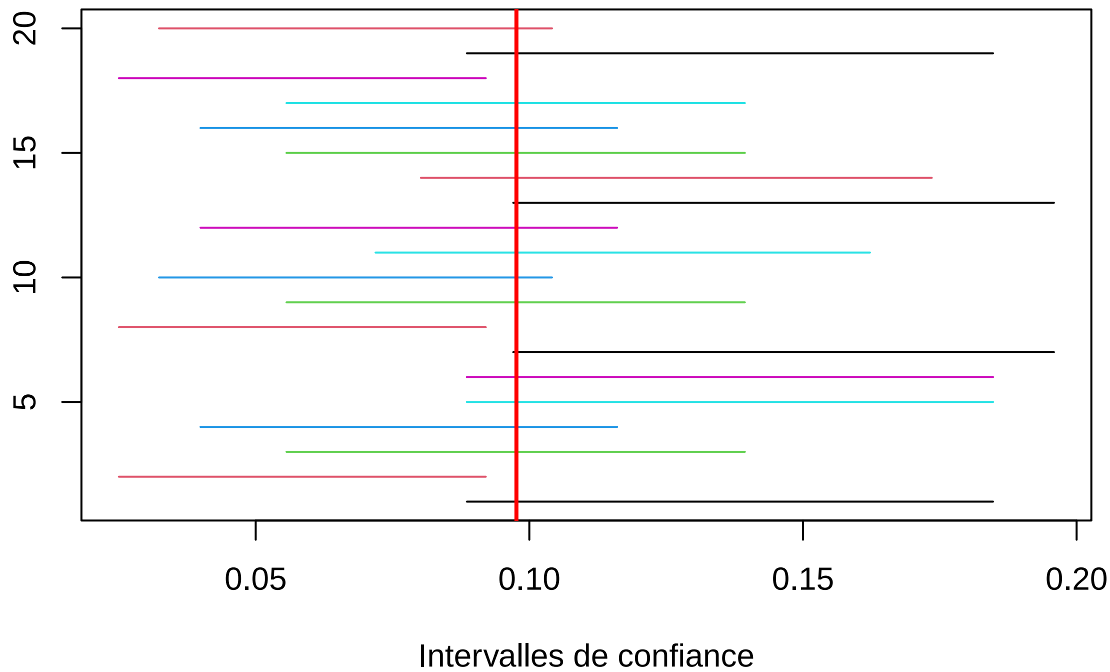
Le vrai taux de chomage (0.1) est presque toujours compris dans l'intervalle, que ce soit celui calcul: celui de la question 5 ou celui de la question 6.

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?

En augmentant n (n = 1000), l'intervalle de confiance se rétrécit. De plus, la proportion de chomeurs calculée dans la question 2 et 4 se rapproche du vértiable taux de chomage. La probabilité que le vrai taux se trouve dans l'intervalle, quant à elle, ne change pas car elle est fixée par notre niveau de confiance (ici: 80 *pourcents).*

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?

```
k <- 20 # nombre des intervalles
ConfInts <- matrix(ncol = k, nrow = 2) # Matrice de 2 lignes et k colonnes
for (i in 1:k){
 # simule le tirage de n éléments
 Sondage17 <- sample(Pop17, n, replace = FALSE)
 # Intervalle de confiance pour la proportion de chômeurs
 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 = "")
# Ajoute la vraie valeur du taux de chomage
abline(v = mean(Pop17), lwd = 2, col = "red")
```



Il y a environ $k\alpha$ intervalles qui ne contiennent pas la vraie valeur du taux de chomage. Avec $k = 20$ et $\alpha = 0.2$, on observe bien qu'en moyenne, il y a 4 intervalle qui ne contiennent pas la vraie valeur du taux de chomage.

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

Comme dit auparavant, la probabilité que le vrai taux de chômage soit dans l'intervalle ne dépend pas de la taille d'échantillon. La seule chose qui change c'est la taille d'intervalles.

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 chomeurs 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 determiner le taux de chomage à 2021. Comparez-la avec l'intervalle de confiance du même seuil pour le taux de chomage à 2017.

```
Pop21 <- rep(0,28900000) #  0 pour la population active (28.9 millions)
Pop21[1:5370000] <- 1 # # 1 pour les gens inscrits à Pole emploi (5.37 millions)
n = 100
Sondage21 <- sample(Pop21,n,replace = FALSE)
tauxSondage = mean(Sondage21) # Moyenne du vecteur Sondage21
# Borne inférieure de l'intervalle de confiance
ICInf = tauxSondage - qnorm(1-alpha/2)*sqrt(tauxSondage*(1-tauxSondage)/n)
```

```r
# Borne supérieure de l'intervalle de confiance
ICSup = tauxSondage + qnorm(1-alpha/2)*sqrt(tauxSondage*(1-tauxSondage)/n)
IC = c(ICInf, ICSup) # vecteur intervalle de confiance
print(IC)
```

```
## [1] 0.1852671 0.2947329
```

L'intervalle de confiance ne contient clairement pas le taux de chômage de 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 different de celui à 2017 (9.8%, i.e., $p = 0.098$). Mathématiquement, on peut formuler les hypothèses du test comme suite:

$$H_0 : p = 0.098$$
$$H_1 : p \neq 0.098.$$

Pour exécutez, on utilise les commandes suivantes (variable `taux` est celui declarée dans la question 2):

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

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(Sondage21) out of n, null probability taux
## X-squared = 21.413, df = 1, p-value = 3.702e-06
## alternative hypothesis: true p is not equal to 0.0976431
## 90 percent confidence interval:
##  0.1728108 0.3218449
## sample estimates:
##    p
## 0.24
```

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

Ici, il faut régarder la p-valeur. Si elle est inférieure à alpha, on rejette H_0. Ici, elle est très proche de 0, alors on rejette H_0 et on conclut que le taux de chômage est significativement different de celui de 2017.