Inférence statistique avec les données de la GCC

Youssef Ait Abdelmalek

24mars 2019

Setup

Charger les packages

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(statsr)
```

```
## Loading required package: BayesFactor
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

Charger le données

```
load("gss.Rdata")
```

Partie 1 : Données

L'ensemble de données utilisé dans cette analyse est l'Enquête sociale générale menée par le National Opinion Research Center. L'ESG contient des données de 1972 à 2012, dans le but de " suivre les changements sociétaux et d'étudier la complexité croissante de la société américaine ". L'objectif est de recueillir des données sur les attitudes, les comportements et les attributs de la société américaine contemporaine et de comparer la société américaine à celle des autres nations. Pour plus d'informations, voir http://www.norc.org/Research/Projects/Pages/general-social-survey.aspx . L'ensemble de données exact utilisé pour cette affectation est un sous-ensemble de l'ensemble de données complet de l'ESG. Toutes les valeurs manquantes ont été codées'NA'.

Partie 2 : Question de recherche

Une question d'actualité aux États-Unis est l'état de l'économie, en particulier le taux de chômage. Une statistique récente indique qu'un homme sur six aux États-Unis est au chômage. Voir : https://twitter.com/NPR/status/773496188209364992/photo/1. L'ensemble de données de l'ESG contient une variable appelée unemp qui enregistre une réponse par oui ou par non à la question suivante posée à chaque répondant : "Au cours des dix dernières années, avez-vous été au chômage et à la recherche d'un emploi pendant plus d'un mois ?" Étant donné que les données de l'ESG couvrent plusieurs décennies, il serait intéressant de comparer les proportions de réponses " oui " d'une année d'enquête récente à une année antérieure pour déterminer s'il y a eu des changements.

Ce rapport comparera la proportion de répondants qui étaient au chômage et à la recherche d'un emploi pendant une période pouvant atteindre un mois au cours de la décennie précédente dans l'année d'enquête la plus récente de 2012, avec la proportion de répondants de la décennie précédente, année d'enquête 2002.

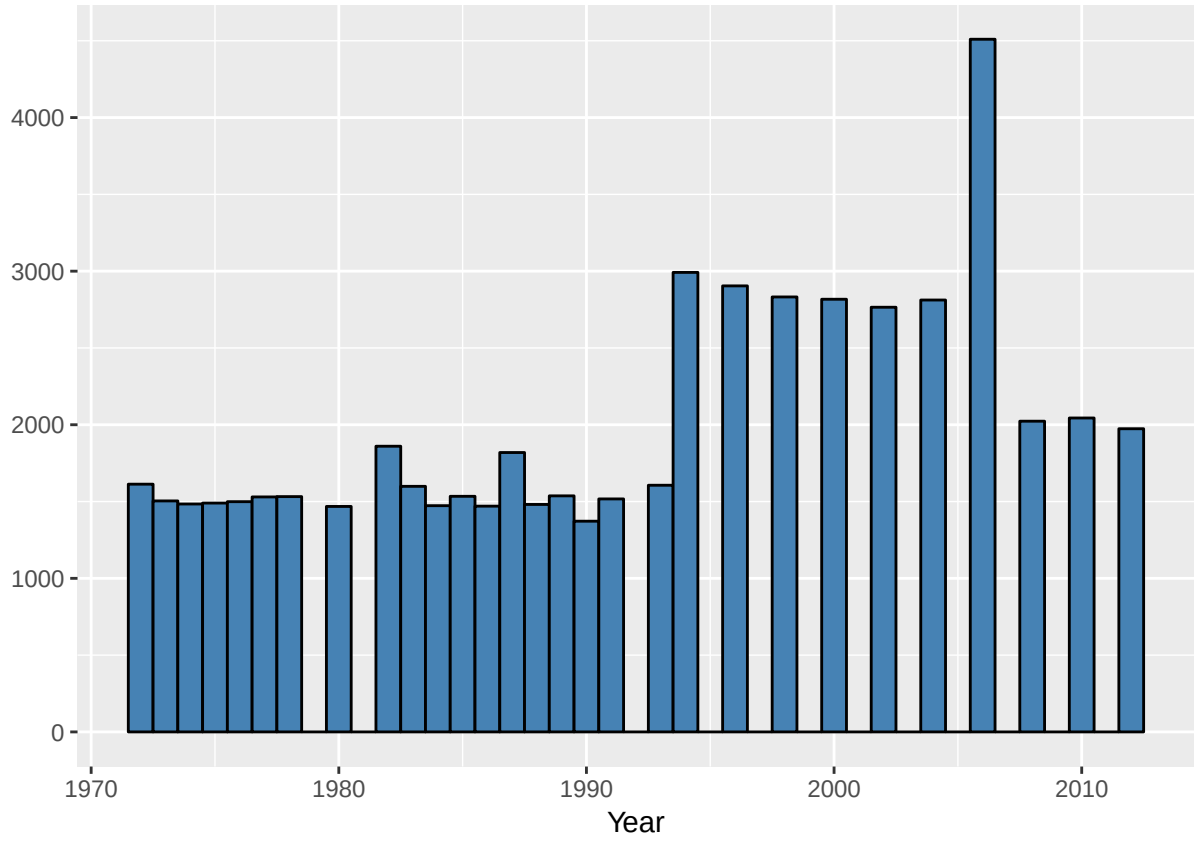
Partie 3 : Analyse exploratoire des données

Avant de choisir la question de recherche finale et les années d'enquête à échantillonner, j'ai généralement observé les données à l'aide de la commande View(gss) et examiné les codes. J'ai décidé de n'utiliser que les variables de l'année et du chômage, mais je voulais mieux comprendre les années de l'enquête. Le code suivant trouvera les valeurs uniques de la variable année et tracera un histogramme de leur fréquence :

```
# trouver les valeurs uniques des années d'enquête dans la variable gss$year
unique(gss$year)
```

```
##  [1] 1972 1973 1974 1975 1976 1977 1978 1980 1982 1983 1984 1985 1986 1987
## [15] 1988 1989 1990 1991 1993 1994 1996 1998 2000 2002 2004 2006 2008 2010
## [29] 2012
```

```
ggplot(gss, aes(x = gss$year)) +
    geom_histogram(binwidth = 1, fill = 'steelblue', col = 'black') +
    xlab('Year') +
    ylab('')
```



J'ai ensuite décidé d'utiliser l'année d'enquête la plus récente, 2012, et de la comparer à la décennie précédente, 2002. Le script suivant réduira le gros fichier de données de l'ESG aux seules variables nécessaires :

```
# réduire l'ensemble de données de l'ESG aux deux seules colonnes nécessaires, " année " et " chômage "
dat <- subset(gss, year == 2002 | year == 2012, select = c(year, sex, unemp))
# Vérifier les dimensions d'un nouvel ensemble de données et afficher un résumé des données.
dim(dat); summary(dat)
```

```
## [1] 4739    3
```

```
##       year          sex         unemp
##  Min.   :2002   Male  :2114   Yes : 757
##  1st Qu.:2002   Female:2625   No  :1490
##  Median :2002                 NA's:2492
##  Mean   :2006
##  3rd Qu.:2012
##  Max.   :2012
```

Il y a pas mal de valeurs de NA dans la variable de chômage, alors je vais les supprimer pour que seules les réponses "Oui" et "Non" restent dans le sous-ensemble de données, et dans le sous-ensemble seulement pour les hommes :

```
# Sous-ensemble à seulement les répondants " Male " et supprimer les répondants de NA
dat <- subset(dat, sex == 'Male', select = c(year, unemp))
# Enlever les valeurs de NA
dat <- na.omit(dat)
# Vérifier les dimensions de l'ensemble de données réduit
dim(dat)
```

```
## [1] 1001    2
```

L'ensemble de données contient maintenant la variable catégorique de l'année avec seulement deux valeurs, " 2002 " et " 2012 ", et unemp, également catégorique avec seulement deux valeurs," Oui " ou " Non ".

Maintenant que j'ai le sous-ensemble exact des données nécessaires, je veux voir un tableau de ces données pour avoir une idée du nombre de réponses " oui " et " non ", ainsi que les proportions :

```
# produire un tableau de données pour voir le nombre réel de réponses " Oui " et " Non ".
table(dat)
```

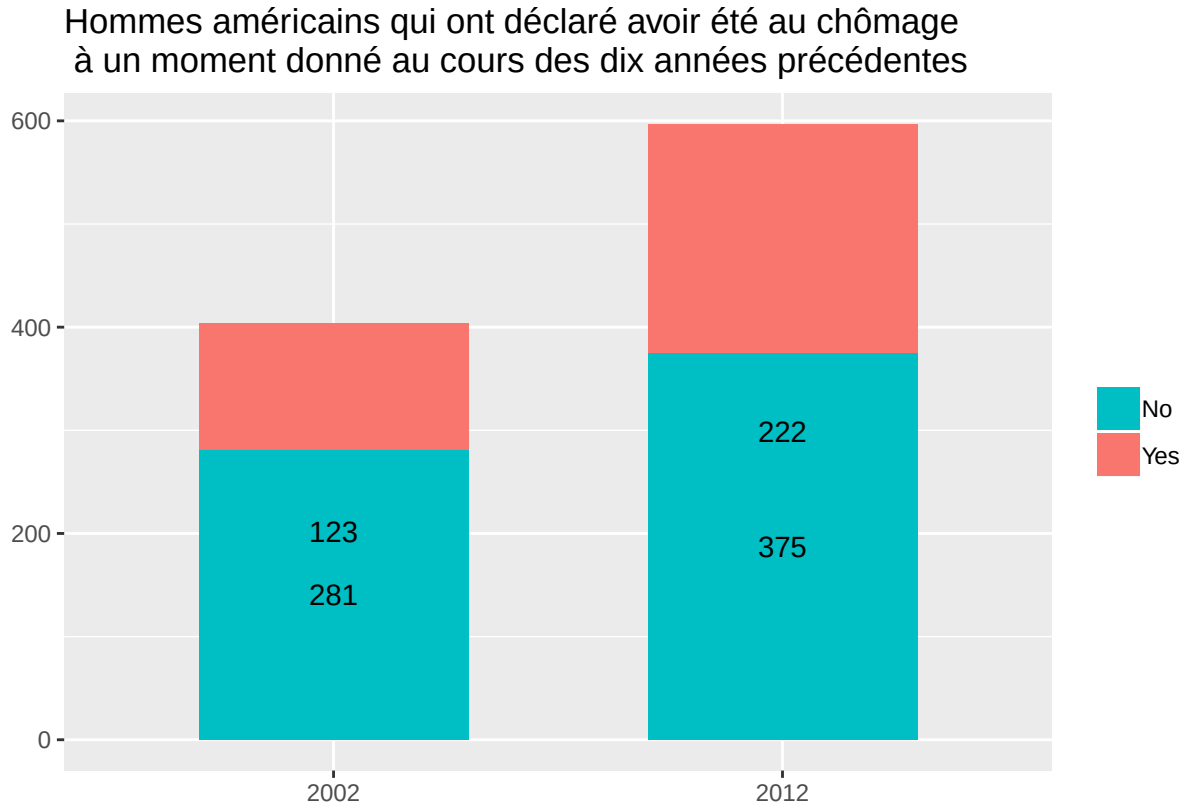
```
##       unemp
## year   Yes  No
##   2002 123 281
##   2012 222 375
```

```
# produire un tableau de fréquence des mêmes données, en arrondissant les proportions à deux décimales
round(prop.table(table(dat),1),2)
```

```
##       unemp
## year    Yes   No
##   2002 0.30 0.70
##   2012 0.37 0.63
```

Enfin, un diagramme à barres des données, montrant le nombre de répondants échantillonnés pour chaque année, avec les proportions de réponses " oui " et " non " :

```
# changer la variable d'année en facteur pour que la fonction plot fonctionne correctement
dat$year  <- as.factor(dat$year)
# plot un bar chart
ggplot(dat, aes(x=year, fill=unemp)) +
    geom_bar(width = 0.6) +
    theme(legend.title=element_blank()) +
    xlab('') +
    ylab('') +
    ggtitle('Hommes américains qui ont déclaré avoir été au chômage \n à un moment donné au cours des d
    guides(fill=guide_legend(reverse=T)) +
    stat_count(aes(label=..count.., y=0.5*..count..), geom = 'text')
```



Pour conclure cette section de l'ADC, le texte qui suit calcule les statistiques d'échantillonnage nécessaires (taille de l'échantillon, nombre de succès et proportions de l'échantillon) :

```
# calculer la taille des échantillons pour les années d'enquête 2002 et 2012
n2002 <- sum(dat$year == 2002)
n2012 <- sum(dat$year == 2012)
# Calculer le nombre de succès (réponses'Oui') dans chaque échantillon.
n2002_Yes <- sum(dat$year == 2002 & dat$unemp == 'Yes')
n2012_Yes <- sum(dat$year == 2012 & dat$unemp == 'Yes')
# calculer les proportions de l'échantillon pour chaque année
phat2002 <- round((n2002_Yes/n2002), 3)
phat2012 <- round((n2012_Yes/n2012), 3)
```

D'après l'analyse exploratoire des données, nous pouvons constater que la proportion de répondants de sexe masculin ayant déclaré avoir été au chômage au cours des dix années précédant l'enquête est passée de $\hat{p}_{2002} = 0,304$, ou 30,4% en 2002, à $\hat{p}_{2012} = 0,372$, ou 37,2%, en 2012. La taille de l'échantillon de 2002 était $n_{2002} = 404$, et le nombre de réponses " oui " était de 123. La taille de l'échantillon de 2012 était de $n_{2012} = 597$, et le nombre de réponses " oui " était de 222. La prochaine section comparera les deux proportions de l'échantillon pour déterminer si ces mesures des données ont une signification statistique.

Partie 4 : Inférence

La dernière partie de ce rapport examinera les deux proportions de l'échantillon pour déterminer ce que l'on peut déduire au sujet des deux populations de 2002 et 2012. Les deux échantillons ont deux variables catégoriques. La variable de groupement est l'année, avec deux valeurs, soit "2002" ou "2012". La variable réponse est la réponse à la question sur le chômage, également avec deux valeurs, "Oui" ou "Non". La réponse "Oui" est considérée comme le "succès". Pour calculer un intervalle de confiance et effectuer un test d'hypothèse, les techniques d'évaluation des variables catégorielles à deux niveaux seront utilisées.

Intervalle de confiance

Tout d'abord, un intervalle de confiance sera calculé pour estimer la différence entre les deux populations à l'aide des proportions de l'échantillon de population. Nous disposons des informations suivantes sur les deux échantillons de populations :

Taille de l'échantillon : $n_{2002} = 404$; $n_{2012} = 597$ Nombre de succès ('Oui') : 123 ; 222 Proportions de l'échantillon : $\hat{p}_{2002} = 0,304$; $\hat{p}_{2012} = 0,372$

La question posée est la suivante : comment les populations masculines des États-Unis de 2002 et 2012 se comparent-elles aux proportions de chômeurs et de personnes à la recherche d'un emploi pendant au moins un mois au cours des dix dernières années ?

Le paramètre d'intérêt est la différence entre les proportions de l'ensemble de la population masculine américaine en 2012 et de l'ensemble de la population masculine américaine en 2002 qui étaient sans emploi et à la recherche d'un emploi pendant au moins un mois au cours des dix années précédentes.

$$p_{2012} - p_{2002}$$

L'estimation ponctuelle est la différence entre les proportions de la population échantillonnée de 2012 et de la population échantillonnée de 2002 qui étaient sans emploi pendant au moins un mois au cours des dix années précédentes.

$$\hat{p}_{2012} - \hat{p}_{2002}$$

La différence entre les proportions sera estimée en calculant un intervalle de confiance, qui est l'estimation ponctuelle plus ou moins la marge d'erreur :

$$\hat{p}_{2012} - \hat{p}_{2002} \pm ME$$

$$\hat{p}_{2012} - \hat{p}_{2002} \pm z * SE_{\hat{p}_{2012} - \hat{p}_{2002}}$$

La proportion de l'échantillon pour 2012 est $\hat{p}_{2012} = 0,372$, et pour 2002 est $\hat{p}_{2002} = 0,304$.

```
# Régler la statistique z à un intervalle de confiance de 95%.
z <- 1.96
```

Pour calculer un intervalle de confiance à 95 %, la statistique z est de 1,96.

L'erreur type pour la différence entre deux proportions est calculée comme suit :

$$SE = \sqrt{(\frac{\hat{p}_{2012}(1 - \hat{p}_{2012})}{n_{2012}} + \frac{\hat{p}_{2002}(1 - \hat{p}_{2002})}{n_{2002}})}$$

```
# calculer le standard error
se <- round(sqrt((phat2012 * (1 - phat2012) / n2012) + (phat2002 * (1 - phat2002) / n2002)), 3)
```

L'erreur type est $SE = 0,03$.

Avant de poursuivre, les conditions d'inférence pour comparer deux proportions indépendantes seront vérifiées :

Premièrement, l'indépendance, c'est-à-dire que les observations doivent être indépendantes au sein de chaque groupe. Il doit y avoir un échantillonnage aléatoire (voir la partie 1) et la condition des 10 % doit être remplie pour les deux échantillons. La taille des échantillons de 597 et 404 est certainement inférieure à dix pour cent de la population masculine américaine (n < 10% de la population, soit plus de 300 millions).

De plus, entre les groupes, les deux groupes doivent être indépendants l'un de l'autre, c'est-à-dire non appariés. Bien qu'il soit théoriquement possible que la même personne ait été interviewée en 2002, puis dix ans plus tard en 2012, étant donné la taille de la population américaine et les techniques d'échantillonnage de l'enquête, cela est très peu probable. Par conséquent, la condition d'indépendance est présumée remplie.

Deuxièmement, la taille de l'échantillon et les conditions de biais, c'est-à-dire que les échantillons doivent répondre à la condition de succès ou d'échec en utilisant les succès et les échecs observés :

$$n_1\hat{p}_1 >= 10; n_1(1-\hat{p}_1) >= 10$$
$$n_2\hat{p}_2 >= 10; n_2(1-\hat{p}_2) >= 10$$

En utilisant les chiffres calculés précédemment : $597 * 0,372 = 221$ succès, soit > 10 ; et $597 * 0,628 = 376$ échecs, soit aussi > 10. $404 * 0,304 = 123$ succès, soit > 10 ; et $404 * 0,696 = 283$ échecs, soit aussi > 10.

Maintenant que les conditions sont remplies, l'intervalle de confiance est calculé comme suit :

$$\hat{p}_{2012} - \hat{p}_{2002} \pm z * SE_{\hat{p}_{2012}-\hat{p}_{2002}}$$

$$(0.372???0.304)ś1.96Ö0.03$$

$$0.068ś0.059$$

```
# calculer les valeurs supérieures et inférieures de l'intervalle de confiance
z <- 1.96

ciupper <- round((phat2012 - phat2002) + (z * se), 2)
cilower <- round((phat2012 - phat2002) - (z * se), 2)
print (ciupper)
```

```
## [1] 0.13
```

```
print (cilower)
```

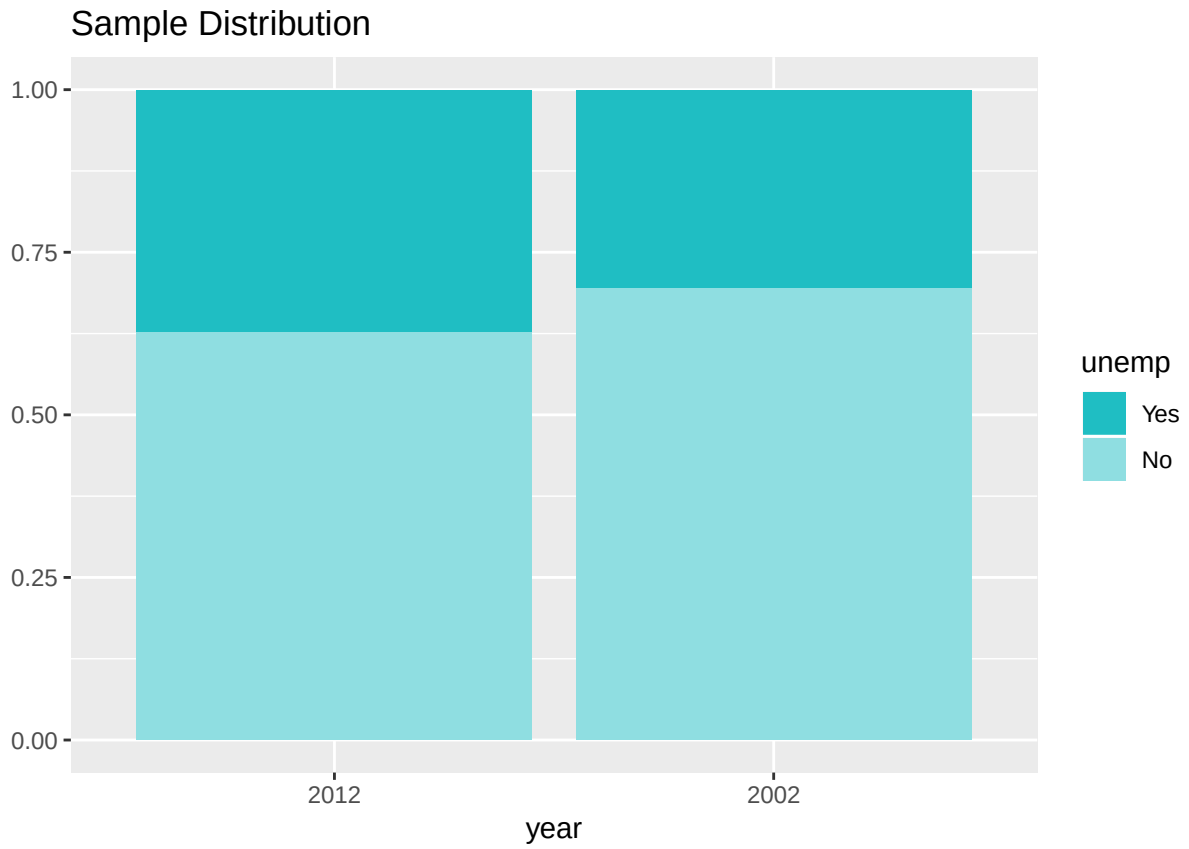
```
## [1] 0.01
```

L'intervalle de confiance est : (0.01, 0.13)

Pour vérifier le calcul manuel de l'intervalle de confiance, le script suivant utilisera le même sous-ensemble de données et la fonction d'inférence du paquet statsr :

```
# utiliser la fonction d'inférence pour calculer l'intervalle de confiance
ci <- inference(unemp, year, dat, type = 'ci', statistic = 'proportion', success = 'Yes', method = 'the
```

```
## Response variable: categorical (2 levels, success: Yes)
## Explanatory variable: categorical (2 levels)
## n_2012 = 597, p_hat_2012 = 0.3719
## n_2002 = 404, p_hat_2002 = 0.3045
## 95% CI (2012 - 2002): (0.0081 , 0.1267)
```




```
# extraire les valeurs CI supérieure et inférieure du résultat de l'inférence
cilower1 <- round(ci$CI[1], 2)
ciupper1 <- round(ci$CI[2], 2)
print (ciupper1)
```

```
## [1] 0.13
```

```
print (cilower1)
```

```
## [1] 0.01
```

La fonction d'inférence produit un graphique de données exploratoires des tailles et des proportions des échantillons pour les années d'enquête 2002 et 2012, qui correspond aux calculs manuels.

L'intervalle de confiance à 95 % de la fonction d'inférence, arrondi à deux décimales près, correspond également au calcul manuel : (0.01, 0.13).

Conclusion sur l'intervalle de confiance : Qu'est-ce que ça veut dire ? Nous sommes convaincus à 95 % que la proportion d'hommes des États-Unis en 2012 qui étaient au chômage et à la recherche d'un emploi pendant au moins un mois au cours de la décennie précédente est de 1 % à 13 % plus élevée que la proportion d'hommes des États-Unis en 2002 qui étaient au chômage et qui cherchaient du travail pendant au moins un mois au cours de la décennie précédente.

Test d'hypothèse

Ensuite, un test d'hypothèse sera effectué avec les mêmes données. L'hypothèse nulle est qu'il n'y a pas de différence dans les proportions de la population de 2002 et 2012 qui étaient sans emploi et à la recherche d'un emploi pendant au moins un mois au cours des dix années précédentes. L'hypothèse alternative est que la

proportion de 2012 est supérieure à celle de 2002.

$$H_0 : \hat{p}_{2012} = \hat{p}_{2002}$$

$$H_A : \hat{p}_{2012} > \hat{p}_{2002}$$

Avant de continuer, les conditions d'exécution d'un test d'hypothèse seront vérifiées. Pour vérifier les conditions d'un test d'hypothèse à deux proportions, on utilise les proportions attendues. Comme ce nombre n'est pas connu, une proportion mise en commun sera calculée et utilisée :

```
# Proportionnalité de l'ensemble caculée
pPool <- round((n2012_Yes + n2002_Yes) / (n2012 + n2002), 3)
print(pPool)
```

```
## [1] 0.345
```

The pooled proportion is $p_{\hat{Pool}} = 0.345$.

As discuss above, because our sample sizes are significantly less than 10% of the male population in the United States, we can assume independence within groups and between groups.

To check the sample size/skew condition, we use the following formula:

$$n_1\hat{p}_{pool} >= 10; n_1(1 - \hat{p}_{pool}) >= 10$$

$$n_2\hat{p}_{pool} >= 10; n_2(1 - \hat{p}_{pool}) >= 10$$

Using the numbers previously calculated: $597 * 0.345 = 206$ successes, which is > 10; and $597 * 0.655 = 391$ failures, which is also > 10. $404 * 0.345 = 139$ successes, which is > 10; and $404 * 0.655 = 265$ failures, which is also > 10.

The 10% condition is met so we can assume the sampling distribution of the difference between the 2012 and 2002 proportions is nearly normal. Therefore, all the test conditions are met.

Next, the standard error will be calculated using p_{pool}:

```
# calculer standard error
se_ht <- round(sqrt((pPool * (1 - pPool) / n2012) + (pPool * (1 - pPool) / n2002)), 3)
print(se_ht)
```

```
## [1] 0.031
```

The standard error is $SE = 0.031$.

Now we are ready to conduct the hypothesis test, at a 5% significance level, evaluating if 2012 males and 2002 males in the United States are equally likely to answer the survey question 'Yes' about whether they were unemployed for at least one month in the prior decade.

The point estimate is:

$$\hat{p}_{2012}???\hat{p}_{2002} = 0.372???0.304 = 0.068$$

```
# régler la valeur nulle à zéro car le test d'hypoothèse suppose que la différence dans les proportions
nullvalue <- 0
```

La valeur nulle est null=0.

La statistique z est calculée comme l'estimation ponctuelle moins la valeur nulle, divisée par l'erreur type :

```r
# calculer l'estimation ponctuelle, qui est la différence dans les proportions observées
pe <- phat2012 - phat2002
# Calculer la statistique z
z_ht <- round(((pe - nullvalue) / se), 3)
print(z_ht)
```

```
## [1] 2.267
```

```r
# Calculer p-value à l'aide de la statistique z ; calculer la valeur de la queue supérieure, car le tes
p_value <- round(pnorm(abs(z_ht), lower.tail = FALSE), 3)
print(p_value)
```

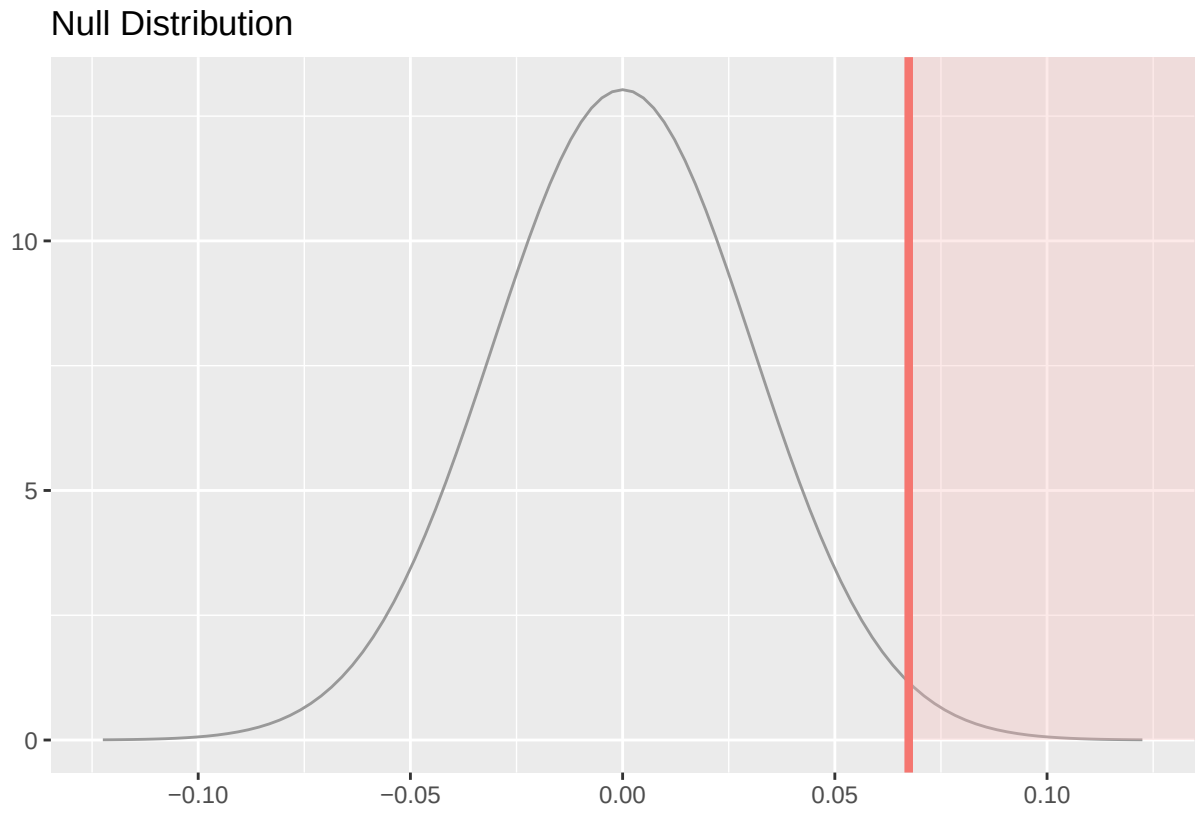
```
## [1] 0.012
```

En utilisant la fonction pnorm, la valeur p de ce z score est 0,012.

Utilisez la fonction d'inférence du paquet statsr pour vérifier le calcul manuel :

```r
# utiliser la fonction d'inférence pour le test d'hypothèse
ht <- inference(unemp, year, dat, type = 'ht', statistic = 'proportion', success = 'Yes', method = 'the
```

```
## Response variable: categorical (2 levels, success: Yes)
## Explanatory variable: categorical (2 levels)
## n_2012 = 597, p_hat_2012 = 0.3719
## n_2002 = 404, p_hat_2002 = 0.3045
## H0: p_2012 =  p_2002
## HA: p_2012 > p_2002
## z = 2.2015
## p_value = 0.0139
```




```
# obtenir p-value du résultat de la fonction d'inférence
ht_p <- round(ht$p, 3)
print(ht_p)
```

```
## [1] 0.014
```

Le graphique produit par la fonction d'inférence montre la distribution nulle du test d'hypothèse avec une ligne à la valeur p. La valeur p de 0,014 est raisonnablement proche de la valeur p calculée manuellement de 0,012.

Conclusion du test d'hypothèse : Qu'est-ce que cela signifie ? En comparant la valeur p calculée de 0,012 à la valeur de signification de 5 %, nous constatons que la valeur p est inférieure à la valeur de signification et concluons qu'il existe des preuves convaincantes que la proportion de la population masculine américaine de 2012 qui a été sans emploi pendant au moins un mois durant la décennie précédente est supérieure à celle de la population féminine américaine de 2002 qui l'a été durant au moins un mois durant la décennie précédente. En raison de la faible valeur p, nous rejetons l'hypothèse nulle qui n'était pas une différence de proportions.

Enfin, l'intervalle de confiance calculé de (0,01, 0,13) n'inclut pas zéro, ce qui correspond au résultat du test d'hypothèse selon lequel le nul (la différence entre les deux proportions est nulle) devrait être rejeté.