Université Internationale de Rabat THE INNOVATIVE UNIVERSITY	**Introduction à l'analyse de données** Mme Nada SBIHI	**Année Universitaire** 2020/2021

Fait par : **EL HANAFI Maha**

Compte Rendu

Lab 6: Inference for Categorical data

```
1  download.file("http://www.openintro.org/stat/data/atheism.RData", destfile = "atheism.RData")
2  load("atheism.RData")
3  view(atheism)
4
```

	nationality	response	year
1	Afghanistan	non-atheist	2012
2	Afghanistan	non-atheist	2012
3	Afghanistan	non-atheist	2012
4	Afghanistan	non-atheist	2012
5	Afghanistan	non-atheist	2012
6	Afghanistan	non-atheist	2012
7	Afghanistan	non-atheist	2012
8	Afghanistan	non-atheist	2012
9	Afghanistan	non-atheist	2012
10	Afghanistan	non-atheist	2012
11	Afghanistan	non-atheist	2012
12	Afghanistan	non-atheist	2012
13	Afghanistan	non-atheist	2012
14	Afghanistan	non-atheist	2012
15	Afghanistan	non-atheist	2012

The survey:

1. These data were taken from a poll so they are based sample statistics. It would not be feasible to know the exact population parameters in this case.
2. The sample observations are independent. The sample size must be < 10% of the population.

The Data:

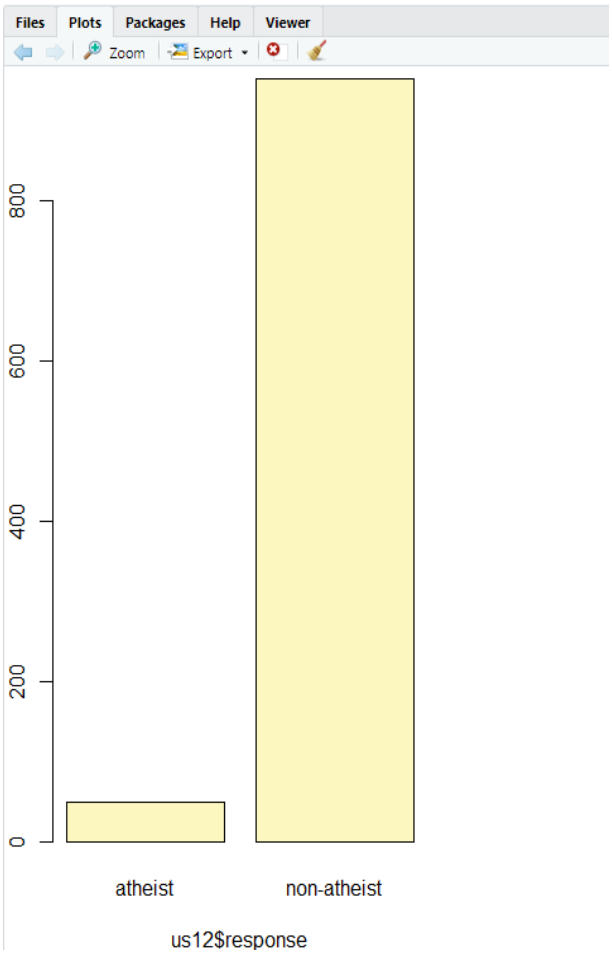
3.
```
14
15  load("more/atheism.RData")
16  names(atheism)
17
```
```
> names(atheism)
[1] "nationality" "response"    "year"
>
```

Per country, % of religious, not religious, atheist and did not respond for the sample size.

4.
```
>
> #Q4
> us12 <- subset(atheism, nationality == "United States" & year == "2012")
> us12ath <- subset(atheism, nationality == "United States" & year == "2012" & response == "atheist")
> nrow(us12ath)/nrow(us12)
[1] 0.0499002
>
```

Inference on proportions:

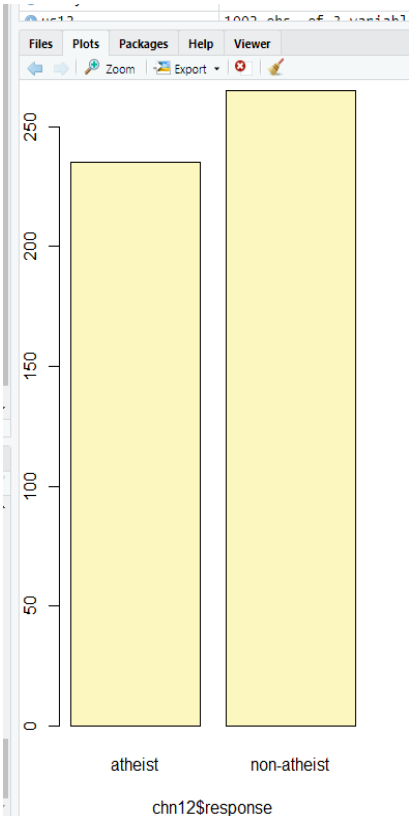
5.
```
>
> #Q5
> inference(us12$response, est = "proportion", type = "ci", method = "theoretical",
+           success = "atheist")
Single proportion -- success: atheist
Summary statistics: p_hat = 0.0499 ;  n = 1002
Check conditions: number of successes = 50 ; number of failures = 952
Standard error = 0.0069
95 % Confidence interval = ( 0.0364 , 0.0634 )
>
```



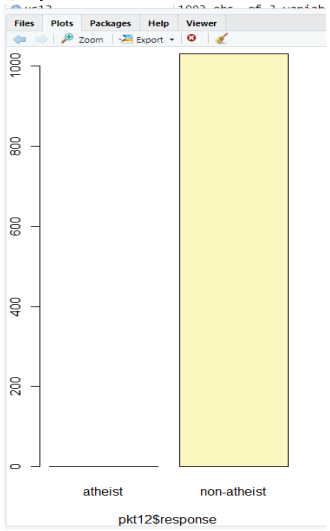
6.
```
>
> #Q6
> SE = 0.0069
> Z_score = 1.96
> ME = SE * Z_score
> ME
[1] 0.013524
>
```

7.

```
> #Q7
> chn12 <- subset(atheism, nationality == "China" & year == "2012")
>
> inference(chn12$response, est = "proportion", type = "ci", method = "theoretical",
+           success = "atheist")
Single proportion -- success: atheist
Summary statistics: p_hat = 0.47 ;  n = 500
Check conditions: number of successes = 235 ; number of failures = 265
Standard error = 0.0223
95 % Confidence interval = ( 0.4263 , 0.5137 )
>
```

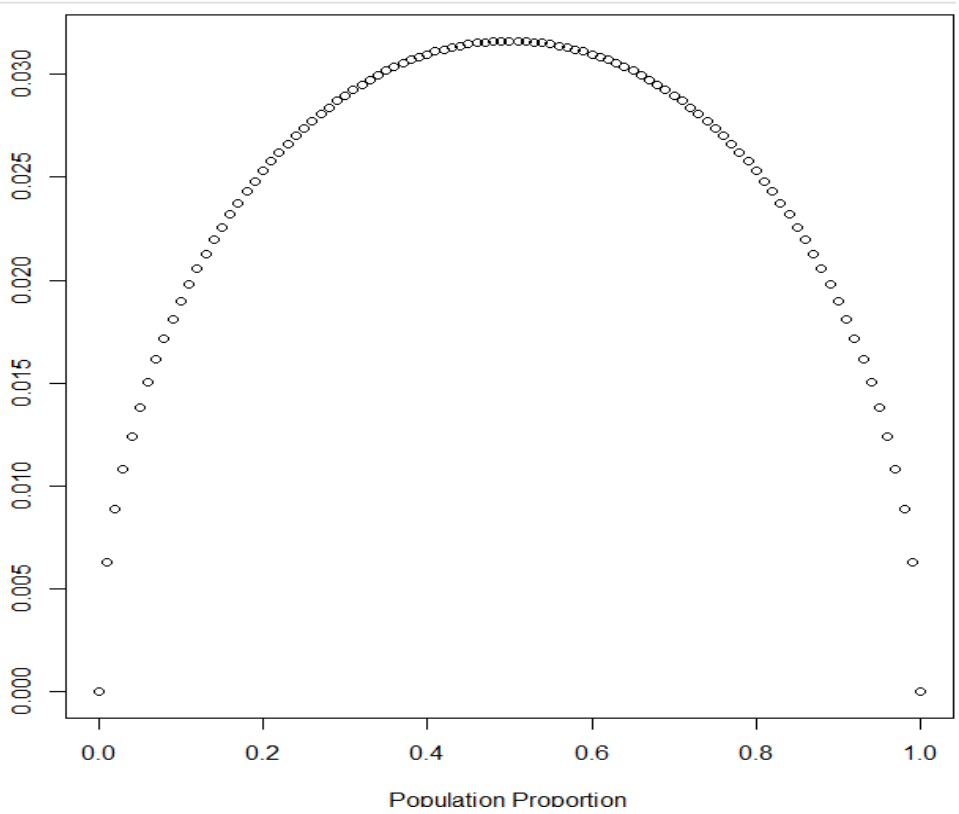



```
> pkt12 <- subset(atheism, nationality == "Afghanistan" & year == "2012")
>
> inference(pkt12$response, est = "proportion", type = "ci", method = "theoretical",
+           success = "atheist")
Single proportion -- success: atheist
Summary statistics: p_hat = 0 ;  n = 1031
Check conditions: number of successes = 0 ; number of failures = 1031
Erreur : There aren't at least 10 successes and 10 failures, use simulation methods instead.
>
```



How does the proportion affect the margin of error?

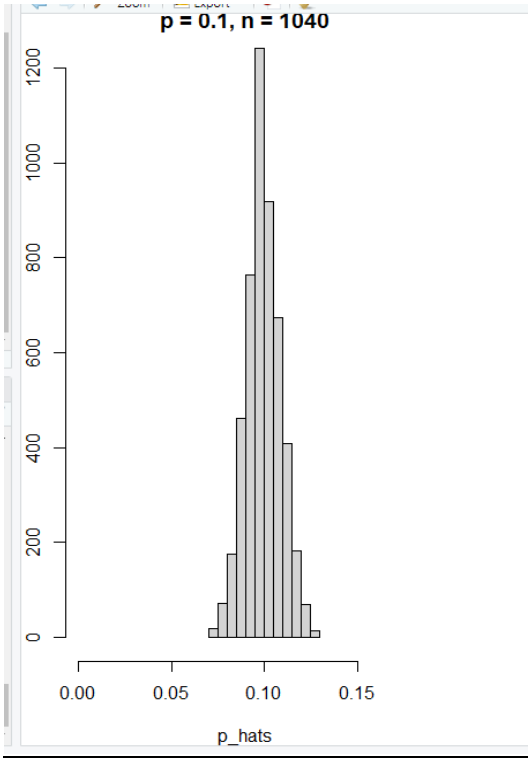
```
>
>
> n <- 1000
> p <- seq(0, 1, 0.01)
> me <- 2 * sqrt(p * (1 - p)/n)
> plot(me ~ p, ylab = "Margin of Error", xlab = "Population Proportion")
>
```



8. It is parabolic.

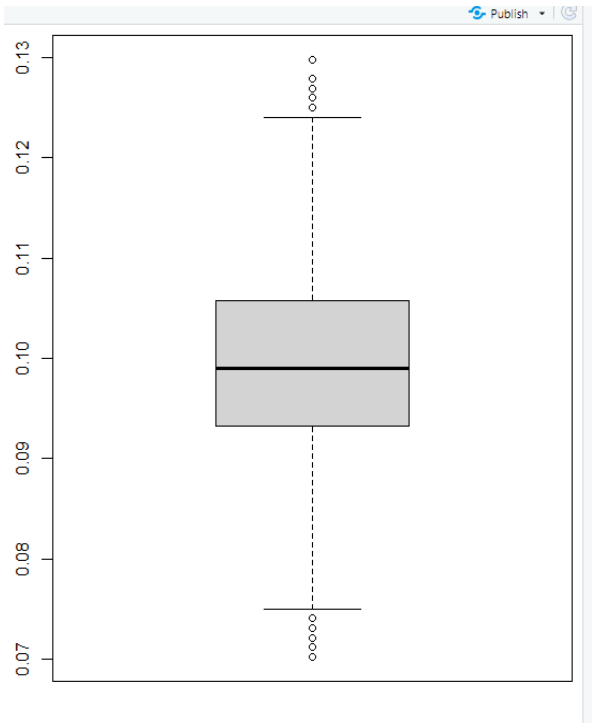
Success-failure condition:

```
>
> p <- 0.1
> n <- 1040
> p_hats <- rep(0, 5000)
>
> for(i in 1:5000){
+   samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
+   p_hats[i] <- sum(samp == "atheist")/n
+ }
>
> hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
> plot(me ~ p, ylab = "Margin of Error", xlab = "Population Proportion")
```



9.

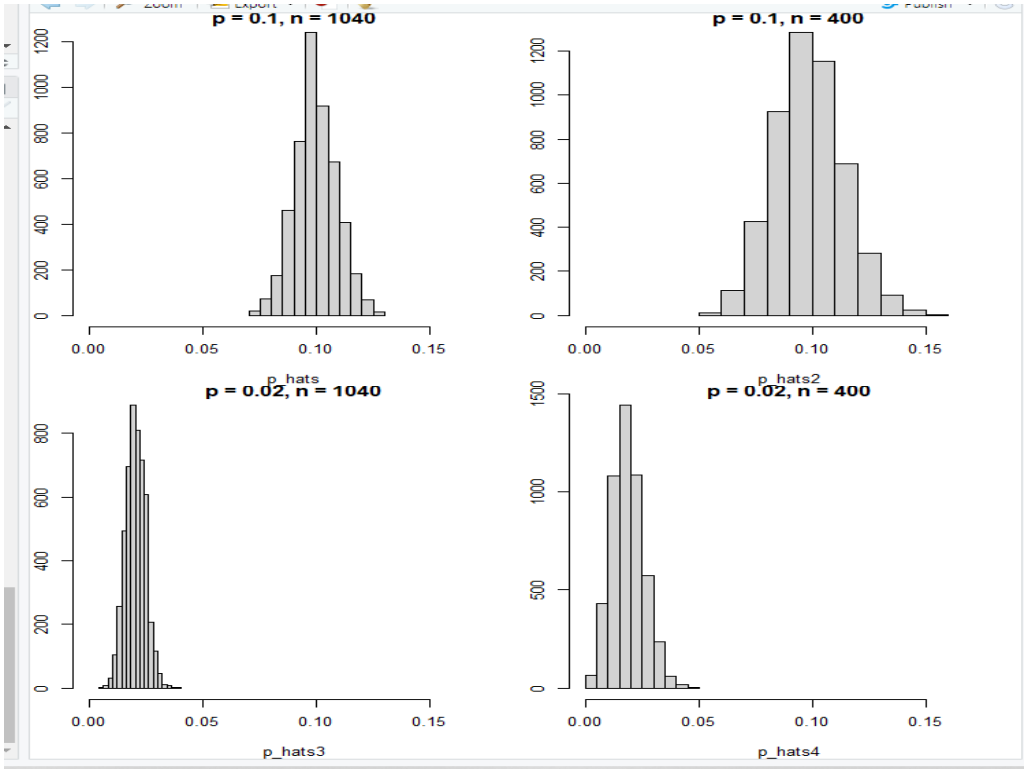
```
> summary(p_hats)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.07019 0.09327 0.09904 0.09969 0.10577 0.12981
> sd(p_hats)
[1] 0.009287382
> boxplot(p_hats,y_lab="p_hats",x_lab="proportions")
>
```



The sampling distribution has a near normal distribution

10.

```
>
> #Q10
> par(mfrow = c(2, 2))
>
> #first histogram
> hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
>
> p <- 0.1
> n <- 400
> p_hats2 <- rep(0, 5000)
>
> for(i in 1:5000){
+   samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
+   p_hats2[i] <- sum(samp == "atheist")/n
+ }
>
> #second histogram
> hist(p_hats2, main = "p = 0.1, n = 400", xlim = c(0, 0.18))
>
> p <- 0.02
> n <- 1040
> p_hats3 <- rep(0, 5000)
>
> for(i in 1:5000){
+   samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
+   p_hats3[i] <- sum(samp == "atheist")/n
+ }
>
> #third histogram
> hist(p_hats3, main = "p = 0.02, n = 1040", xlim = c(0, 0.18))
>
> p <- 0.02
> n <- 400
> p_hats4 <- rep(0, 5000)
>
> for(i in 1:5000){
+   samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
+   p_hats4[i] <- sum(samp == "atheist")/n
+ }
>
> #fourth histogram
> hist(p_hats4, main = "p = 0.02, n = 400", xlim = c(0, 0.18))
>
```



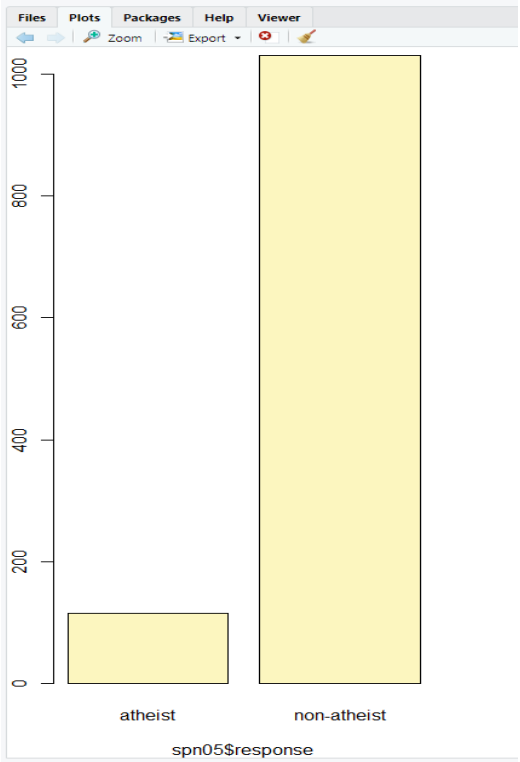
11. They both have a normal distribution with almost similar spreads.

On your own:

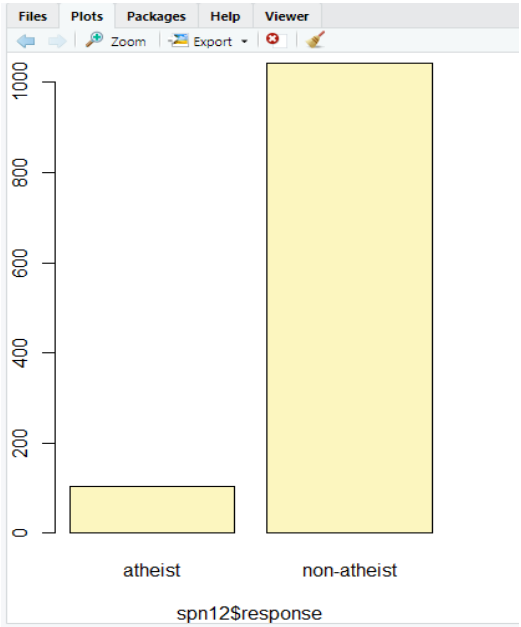
1.

a.

```
>
> #On your own
> spn05 <- subset(atheism, nationality == "Spain" & year == "2005")
> spn12 <- subset(atheism, nationality == "Spain" & year == "2012")
>
> inference(spn05$response, est = "proportion", type = "ci", method = "theoretical",
+           success = "atheist")
Single proportion -- success: atheist
Summary statistics: p_hat = 0.1003 ;  n = 1146
Check conditions: number of successes = 115 ; number of failures = 1031
Standard error = 0.0089
95 % Confidence interval = ( 0.083 , 0.1177 )
>
```



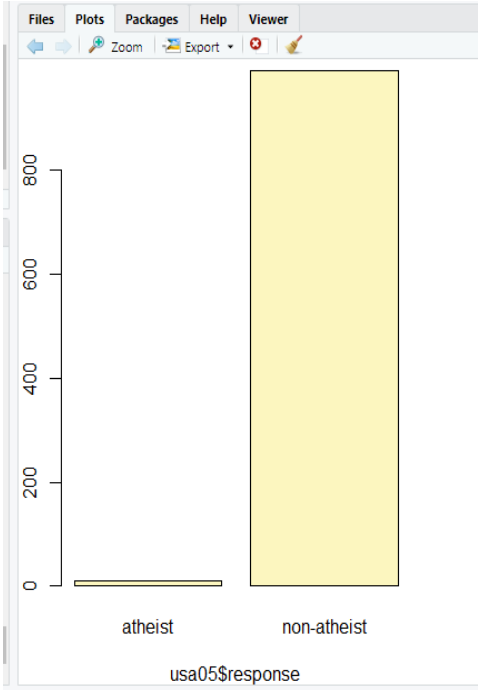

```
> inference(spn12$response, est = "proportion", type = "ci", method = "theoretical",
+           success = "atheist")
Single proportion -- success: atheist
Summary statistics: p_hat = 0.09 ;  n = 1145
Check conditions: number of successes = 103 ; number of failures = 1042
Standard error = 0.0085
95 % Confidence interval = ( 0.0734 , 0.1065 )
>
```




```
95 % Confidence interval = ( 0.0734 , 0.1065 )
> #standard
> p_spn05 = 0.1003
> n_spn05 = 1146
> p_spn12 = 0.09
> n_spn12 = 1145
>
> PE_spn = p_spn12 - p_spn05
>
> SE_spn = sqrt((p_spn05*(1-p_spn05)/n_spn05)+(p_spn12*(1-p_spn12)/n_spn12))
> SE_spn
[1] 0.01225854
>
```

b.

```
> usa05 <- subset(atheism, nationality == "United States" & year == "2005")
> usa12 <- subset(atheism, nationality == "United States" & year == "2012")
>
> inference(usa05$response, est = "proportion", type = "ci", method = "theoretical",
+           success = "atheist")
Single proportion -- success: atheist
Summary statistics: p_hat = 0.01 ;  n = 1002
Check conditions: number of successes = 10 ; number of failures = 992
Standard error = 0.0031
95 % Confidence interval = ( 0.0038 , 0.0161 )
>
```



3.

```
95 % Confidence interval = ( 0.0368 , 0.0641 )
> #e3
> SE = 0.01/1.96
> p =0.018
> n = (p*(1-p))/SE^2
> ceiling(n)
[1] 680
>
```