Capsule 12: Tester une hypothèse

Janie Coulombe

2022-08-01

J'utilise dans cet aide-mémoire un jeu de données [1] sur un essai randomisé sur l'ocytocine pour la prévention des hémorragies post partum, que j'ai importé et enregistré dans `data2`. Voici les variables contenues dans le jeu de données (la fonction `colnames` nous renvoie le nom des colonnes, c.-à-d. les variables):

```
colnames(data2)
```

```
##   [1] "patientID"                   "inicialesm"                 
##   [3] "inicialesp"                  "fecha"                      
##   [5] "dar_a_luz"                   "parto"                      
##   [7] "enfermedad"                  "enfermedad_especifique"     
##   [9] "elegible"                    "participar"                 
##  [11] "consentimiento"              "edad"                       
##  [13] "estudios"                    "estadocivil"                
##  [15] "numembarazos"                "numhijos"                   
##  [17] "fecha_ingreso"               "edadgest"                   
##  [19] "hemorragia"                  "ninguna"                    
##  [21] "hipertension_cronica"        "pre_eclampsia"              
##  [23] "diabetes"                    "epilepsia"                  
##  [25] "asma"                        "otro"                       
##  [27] "otro_espec"                  "Nivel_hb"                   
##  [29] "presion_sis"                 "presion_dia"                
##  [31] "frecuencia"                  "ID_2"                       
##  [33] "inicialesm_2"                "inicialesp_2"               
##  [35] "fecha_2"                     "cesarea"                    
##  [37] "time_nacimiento"             "neonato_unico"              
##  [39] "coloco_IV"                   "epidural"                   
##  [41] "induccion"                   "induccion_espec"            
##  [43] "estimulacion"                "estimulacion_espec"         
##  [45] "anti_hipertensivos"          "anti_hipertensivos_espec"   
##  [47] "antibioticos"                "antibioticos_espec"         
##  [49] "magnesio"                    "magnesio_espec"             
##  [51] "otro_medicamento"            "otro_medicamento_espec"     
##  [53] "episiotomia"                 "paquete_AR"                 
##  [55] "studygrp_AR"                 "hora_profilaxis_AR"         
##  [57] "problema_medicamento_AR"     "problema_medicamento_espec_AR"
##  [59] "cantidad_IV_AR"              "tamano_AR"                  
##  [61] "tamano_espec_AR"             "velocidad_espec_AR"         
##  [63] "traccion_cordon"             "masajeuterino"              
##  [65] "placenta_time"               "extraccionmanual"           
##  [67] "hora_15"                     "presion_sist_15"            
##  [69] "presion_diast_15"            "frecuencia_15"              
##  [71] "hora_30"                     "presion_sist_30"            
##  [73] "presion_diast_30"            "frecuencia_30"              
##  [75] "sangre_30"                   "hora_45"                    
##  [77] "presion_sist_45"             "presion_diast_45"           
##  [79] "frecuencia_45"               "hora_60"                    
##  [81] "presion_sist_60"             "presion_diast_60"           
##  [83] "frecuencia_60"               "sangre_60"                  
##  [85] "problemas_medicion"          "problemas_medicion_expl"    
##  [87] "hpp"                         "perdida_sangre_hpp"         
##  [89] "hora_hpp"                    "utero_contraido"            
##  [91] "causa_HPP"                   "perdida_sangre_total"       
##  [93] "hora_detuvo"                 "segundaro_efecto_AR"        
##  [95] "addl_comments"               "sutura"                     
##  [97] "compresion"                  "exploracion"                
##  [99] "uterotonicos"                "medicamento1"               
## [101] "via1"                        "medicamento2"               
## [103] "via2"                        "histerectomia"              
## [105] "otra_cirugia"                "otra_cirugia_txt"           
## [107] "transfusion"                 "transfusion_txt"            
## [109] "unidades"                    "expansores"                 
## [111] "expansores_espec"            "otra"                       
## [113] "otra_txt"                    "estado_mujer"               
## [115] "fecha_fallecio"              "time_venoclisis"            
## [117] "volumen_electrolitos"        "time_hb"                    
## [119] "nivel_hb_4"                  "totalbld_ARGoxy"            
## [121] "totalbld500"                 "totalbld750"                
## [123] "totalbld1000"                "totalbld1500"               
## [125] "SI_baseline"                 "SI_15min"                   
## [127] "SI_30min"                    "SI_45min"                   
## [129] "SI_60min"                    "gestage"                    
## [131] "prehbunder11"                "indoraug"                   
## [133] "tpd"                         "Hbdiff"                     
## [135] "HBdrop2ormore"               "Hbdrop2plusORbldtrans"      
## [137] "deliverytime"                "time_diagnosis"             
## [139] "time_prophylaxis"            "timebldcontrolled"          
## [141] "SI15min_threshold0.9"        "SI30min_threshold0.9"       
## [143] "SI45min_threshold0.9"        "SI60min_threshold0.9"       
## [145] "SI_delta_baseto15min"        "SI_delta_baseto30min"       
## [147] "SI_delta_baseto45min"        "SI_delta_baseto60min"
```

```
summary(data2)
```

```
##    patientID     inicialesm         inicialesp            fecha           
##  Min.   :3001   Length:543         Length:543         Min.   :2016-12-01  
##  1st Qu.:3140   Class :character   Class :character   1st Qu.:2017-02-16  
##  Median :3277   Mode  :character   Mode  :character   Median :2017-05-04  
##  Mean   :3276                                         Mean   :2017-04-29  
##  3rd Qu.:3412                                         3rd Qu.:2017-07-04  
##  Max.   :3550                                         Max.   :2017-09-30  
##                                                                           
##    dar_a_luz     parto     enfermedad enfermedad_especifique    elegible
##  Min.   :1   Min.   :0   Min.   :0    Length:543             Min.   :1  
##  1st Qu.:1   1st Qu.:0   1st Qu.:0    Class :character       1st Qu.:1  
##  Median :1   Median :0   Median :0    Mode  :character       Median :1  
##  Mean   :1   Mean   :0   Mean   :0                           Mean   :1  
##  3rd Qu.:1   3rd Qu.:0   3rd Qu.:0                           3rd Qu.:1  
##  Max.   :1   Max.   :0   Max.   :0                           Max.   :1  
##                                                                           
##    participar consentimiento      edad          estudios      estadocivil  
##  Min.   :1    Min.   :1      Min.   :15.00   Min.   :1.000   Min.   :1.000  
##  1st Qu.:1    1st Qu.:1      1st Qu.:20.00   1st Qu.:2.000   1st Qu.:1.000  
##  Median :1    Median :1      Median :23.00   Median :3.000   Median :2.000  
##  Mean   :1    Mean   :1      Mean   :24.34   Mean   :2.702   Mean   :1.556  
##  3rd Qu.:1    3rd Qu.:1      3rd Qu.:28.00   3rd Qu.:3.000   3rd Qu.:2.000  
##  Max.   :1    Max.   :1      Max.   :42.00   Max.   :5.000   Max.   :3.000  
##                                                                              
##   numembarazos       numhijos      fecha_ingreso           edadgest    
##  Min.   : 1.000   Min.   : 0.000   Min.   :2016-11-25   Min.   :25.60  
##  1st Qu.: 1.000   1st Qu.: 0.000   1st Qu.:2017-02-16   1st Qu.:38.10  
##  Median : 2.000   Median : 1.000   Median :2017-05-04   Median :39.10  
##  Mean   : 2.343   Mean   : 1.239   Mean   :2017-04-29   Mean   :38.68  
##  3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.:2017-07-02   3rd Qu.:40.00  
##  Max.   :12.000   Max.   :11.000   Max.   :2017-09-30   Max.   :42.10  
##                                                                         
##    hemorragia        ninguna       hipertension_cronica pre_eclampsia    
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000000     Min.   :0.00000  
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.000000     1st Qu.:0.00000  
##  Median :0.0000   Median :1.0000   Median :0.000000     Median :0.00000  
##  Mean   :0.3112   Mean   :0.9208   Mean   :0.003683     Mean   :0.01289  
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.000000     3rd Qu.:0.00000  
##  Max.   :2.0000   Max.   :1.0000   Max.   :1.000000     Max.   :1.00000  
##                                                                          
##     diabetes         epilepsia             asma        otro        
##  Min.   :0.00000   Min.   :0.000000   Min.   :0   Min.   :0.00000  
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0   1st Qu.:0.00000  
##  Median :0.00000   Median :0.000000   Median :0   Median :0.00000  
##  Mean   :0.01657   Mean   :0.003683   Mean   :0   Mean   :0.05525  
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0   3rd Qu.:0.00000  
##  Max.   :1.00000   Max.   :1.000000   Max.   :0   Max.   :1.00000  
##                                                                    
##   otro_espec           Nivel_hb      presion_sis     presion_dia    
##  Length:543         Min.   : 7.10   Min.   : 73.0   Min.   : 43.00  
##  Class :character   1st Qu.:10.80   1st Qu.:106.0   1st Qu.: 68.00  
##  Mode  :character   Median :11.85   Median :115.0   Median : 75.00  
##                     Mean   :11.75   Mean   :115.4   Mean   : 75.53  
##                     3rd Qu.:12.70   3rd Qu.:124.0   3rd Qu.: 82.00  
##                     Max.   :16.80   Max.   :173.0   Max.   :118.00  
##                     NA's   :5       NA's   :4       NA's   :4       
##    frecuencia          ID_2      inicialesm_2       inicialesp_2      
##  Min.   : 53.00   Min.   :3001   Length:543         Length:543        
##  1st Qu.: 75.00   1st Qu.:3139   Class :character   Class :character  
##  Median : 84.00   Median :3276   Mode  :character   Mode  :character  
##  Mean   : 84.32   Mean   :3275                                        
##  3rd Qu.: 93.00   3rd Qu.:3411                                        
##  Max.   :144.00   Max.   :3550                                        
##  NA's   :4        NA's   :3                                           
##     fecha_2              cesarea       time_nacimiento              
##  Min.   :2016-12-01   Min.   :0.0000   Min.   :2016-12-02 02:05:00  
##  1st Qu.:2017-02-16   1st Qu.:0.0000   1st Qu.:2017-02-15 06:25:30  
##  Median :2017-05-03   Median :0.0000   Median :2017-04-30 04:27:30  
##  Mean   :2017-04-29   Mean   :0.1111   Mean   :2017-04-28 13:49:12  
##  3rd Qu.:2017-07-03   3rd Qu.:0.0000   3rd Qu.:2017-07-02 21:21:00  
##  Max.   :2017-09-30   Max.   :1.0000   Max.   :2017-09-30 21:50:00  
##  NA's   :3            NA's   :3        NA's   :63                   
##  neonato_unico      coloco_IV     epidural    induccion      
##  Min.   :0.0000   Min.   :1    Min.   :0    Min.   :0.00000  
##  1st Qu.:1.0000   1st Qu.:1    1st Qu.:0    1st Qu.:0.00000  
##  Median :1.0000   Median :1    Median :0    Median :0.00000  
##  Mean   :0.9896   Mean   :1    Mean   :0    Mean   :0.08542  
##  3rd Qu.:1.0000   3rd Qu.:1    3rd Qu.:0    3rd Qu.:0.00000  
##  Max.   :1.0000   Max.   :1    Max.   :0    Max.   :1.00000  
##  NA's   :63       NA's   :63   NA's   :63   NA's   :63       
##  induccion_espec     estimulacion    estimulacion_espec anti_hipertensivos
##  Length:543         Min.   :0.0000   Length:543         Min.   :0.00000   
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.:0.00000   
##  Mode  :character   Median :0.0000   Mode  :character   Median :0.00000   
##                     Mean   :0.1396                      Mean   :0.02083   
##                     3rd Qu.:0.0000                      3rd Qu.:0.00000   
##                     Max.   :1.0000                      Max.   :1.00000   
##                     NA's   :63                          NA's   :63        
##  anti_hipertensivos_espec  antibioticos     antibioticos_espec
##  Length:543               Min.   :0.00000   Length:543        
##  Class :character         1st Qu.:0.00000   Class :character  
##  Mode  :character         Median :0.00000   Mode  :character  
##                           Mean   :0.06458                     
##                           3rd Qu.:0.00000                     
##                           Max.   :1.00000                     
##                           NA's   :63                          
##     magnesio       magnesio_espec     otro_medicamento  otro_medicamento_espec
##  Min.   :0.00000   Length:543         Min.   :0.00000   Length:543            
##  1st Qu.:0.00000   Class :character   1st Qu.:0.00000   Class :character      
##  Median :0.00000   Mode  :character   Median :0.00000   Mode  :character      
##  Mean   :0.01458                      Mean   :0.01458                         
##  3rd Qu.:0.00000                      3rd Qu.:0.00000                         
##  Max.   :1.00000                      Max.   :1.00000                         
##  NA's   :63                           NA's   :63                              
##   episiotomia       paquete_AR     studygrp_AR    hora_profilaxis_AR
##  Min.   :0.0000   Min.   :  1.0   Min.   :1.000   Length:543        
##  1st Qu.:0.0000   1st Qu.:121.8   1st Qu.:1.000   Class1:hms        
##  Median :0.0000   Median :241.5   Median :2.000   Class2:difftime   
##  Mean   :0.4479   Mean   :241.8   Mean   :1.502   Mode  :numeric    
##  3rd Qu.:1.0000   3rd Qu.:362.2   3rd Qu.:2.000                     
##  Max.   :1.0000   Max.   :482.0   Max.   :2.000                     
##  NA's   :63       NA's   :63      NA's   :63                        
##  problema_medicamento_AR problema_medicamento_espec_AR cantidad_IV_AR
##  Min.   :0.00000         Length:543                    Min.   :1     
##  1st Qu.:0.00000         Class :character              1st Qu.:1     
##  Median :0.00000         Mode  :character              Median :1     
##  Mean   :0.00417                                       Mean   :1     
##  3rd Qu.:0.00000                                       3rd Qu.:1     
##  Max.   :1.00000                                       Max.   :1     
##  NA's   :63                                            NA's   :63    
##    tamano_AR    tamano_espec_AR    velocidad_espec_AR traccion_cordon 
##  Min.   :1.00   Length:543         Min.   : 10.0      Min.   :0.0000  
##  1st Qu.:1.00   Class :character   1st Qu.: 40.0      1st Qu.:1.0000  
##  Median :1.00   Mode  :character   Median : 40.0      Median :1.0000  
##  Mean   :1.43                      Mean   : 40.2      Mean   :0.9708  
##  3rd Qu.:2.00                      3rd Qu.: 40.0      3rd Qu.:1.0000  
##  Max.   :2.00                      Max.   :100.0      Max.   :1.0000  
##  NA's   :64                        NA's   :67         NA's   :63      
##  masajeuterino    placenta_time                 extraccionmanual 
##  Min.   :0.0000   Min.   :2016-12-02 02:09:00   Min.   :0.00000  
##  1st Qu.:0.0000   1st Qu.:2017-02-15 03:25:30   1st Qu.:0.00000  
##  Median :0.0000   Median :2017-04-30 13:14:00   Median :0.00000  
##  Mean   :0.3583   Mean   :2017-04-28 15:38:01   Mean   :0.00625  
##  3rd Qu.:1.0000   3rd Qu.:2017-07-03 01:40:00   3rd Qu.:0.00000  
##  Max.   :1.0000   Max.   :2017-09-30 22:00:00   Max.   :1.00000  
##  NA's   :63       NA's   :64                    NA's   :63       
##    hora_15          presion_sist_15 presion_diast_15 frecuencia_15   
##  Length:543        Min.   : 58.0   Min.   : 38.00   Min.   : 54.00  
##  Class1:hms        1st Qu.:100.0   1st Qu.: 60.00   1st Qu.: 76.75  
##  Class2:difftime   Median :108.0   Median : 68.00   Median : 87.00  
##  Mode  :numeric    Mean   :108.7   Mean   : 68.96   Mean   : 88.17  
##                    3rd Qu.:117.0   3rd Qu.: 76.00   3rd Qu.: 98.00  
##                    Max.   :164.0   Max.   :114.00   Max.   :154.00  
##                    NA's   :63      NA's   :63       NA's   :63      
##    hora_30          presion_sist_30 presion_diast_30 frecuencia_30   
##  Length:543        Min.   : 40.0   Min.   : 20.00   Min.   : 51.00  
##  Class1:hms        1st Qu.:100.0   1st Qu.: 62.00   1st Qu.: 74.00  
##  Class2:difftime   Median :107.0   Median : 69.00   Median : 83.00  
##  Mode  :numeric    Mean   :107.8   Mean   : 69.28   Mean   : 85.45  
##                    3rd Qu.:115.0   3rd Qu.: 76.00   3rd Qu.: 95.00  
##                    Max.   :166.0   Max.   :111.00   Max.   :153.00  
##                    NA's   :65      NA's   :65       NA's   :65      
##    sangre_30        hora_45         presion_sist_45 presion_diast_45
##  Min.   :   0.0   Length:543        Min.   : 50.0   Min.   : 30.00  
##  1st Qu.: 100.0   Class1:hms        1st Qu.:100.0   1st Qu.: 62.00  
##  Median : 200.0   Class2:difftime   Median :108.0   Median : 69.00  
##  Mean   : 311.3   Mode  :numeric    Mean   :108.2   Mean   : 69.22  
##  3rd Qu.: 400.0                     3rd Qu.:115.0   3rd Qu.: 76.00  
##  Max.   :1700.0                     Max.   :184.0   Max.   :118.00  
##  NA's   :65                         NA's   :66      NA's   :66      
##  frecuencia_45      hora_60         presion_sist_60 presion_diast_60
##  Min.   : 48.00   Length:543        Min.   : 60.0   Min.   : 38.00  
##  1st Qu.: 74.00   Class1:hms        1st Qu.:100.0   1st Qu.: 62.00  
##  Median : 83.00   Class2:difftime   Median :108.0   Median : 70.00  
##  Mean   : 85.07   Mode  :numeric    Mean   :108.4   Mean   : 69.14  
##  3rd Qu.: 96.00                     3rd Qu.:116.0   3rd Qu.: 76.00  
##  Max.   :150.00                     Max.   :164.0   Max.   :102.00  
##  NA's   :66                         NA's   :66      NA's   :66      
##  frecuencia_60      sangre_60      problemas_medicion problemas_medicion_expl
##  Min.   : 46.00   Min.   :  20.0   Min.   :0.00000    Length:543             
##  1st Qu.: 73.00   1st Qu.: 150.0   1st Qu.:0.00000    Class :character       
##  Median : 82.00   Median : 300.0   Median :0.00000    Mode  :character       
##  Mean   : 84.56   Mean   : 382.7   Mean   :0.01875                           
##  3rd Qu.: 95.00   3rd Qu.: 450.0   3rd Qu.:0.00000                           
##  Max.   :150.00   Max.   :2000.0   Max.   :1.00000                           
##  NA's   :66       NA's   :67       NA's   :63                                
##       hpp         perdida_sangre_hpp   hora_hpp        utero_contraido
##  Min.   :0.0000   Min.   : 500.0     Length:543        Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.: 500.0     Class1:hms        1st Qu.:3.000  
##  Median :0.0000   Median : 600.0     Class2:difftime   Median :4.000  
##  Mean   :0.2188   Mean   : 700.9     Mode  :numeric    Mean   :3.865  
##  3rd Qu.:0.0000   3rd Qu.: 800.0                       3rd Qu.:5.000  
##  Max.   :1.0000   Max.   :1700.0                       Max.   :5.000  
##  NA's   :63       NA's   :438                          NA's   :439    
##    causa_HPP     perdida_sangre_total hora_detuvo       segundaro_efecto_AR
##  Min.   :1.000   Min.   :  20.0       Length:543        Min.   :0          
##  1st Qu.:1.000   1st Qu.: 150.0       Class1:hms        1st Qu.:0          
##  Median :4.000   Median : 300.0       Class2:difftime   Median :0          
##  Mean   :2.924   Mean   : 384.9       Mode  :numeric    Mean   :0          
##  3rd Qu.:4.000   3rd Qu.: 450.0                         3rd Qu.:0          
##  Max.   :6.000   Max.   :2000.0                         Max.   :0          
##  NA's   :438     NA's   :66                             NA's   :65         
##  addl_comments          sutura      compresion       exploracion    
##  Length:543         Min.   :0.0   Min.   :0.00000   Min.   :0.0000  
##  Class :character   1st Qu.:0.0   1st Qu.:0.00000   1st Qu.:0.0000  
##  Mode  :character   Median :0.0   Median :0.00000   Median :0.0000  
##                     Mean   :0.4   Mean   :0.00625   Mean   :0.0125  
##                     3rd Qu.:1.0   3rd Qu.:0.00000   3rd Qu.:0.0000  
##                     Max.   :1.0   Max.   :1.00000   Max.   :1.0000  
##                     NA's   :63    NA's   :63        NA's   :63      
##   uterotonicos     medicamento1           via1           medicamento2      
##  Min.   :0.00000   Length:543         Length:543         Length:543        
##  1st Qu.:0.00000   Class :character   Class :character   Class :character  
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :0.08958                                                           
##  3rd Qu.:0.00000                                                           
##  Max.   :1.00000                                                           
##  NA's   :63                                                                
##      via2           histerectomia  otra_cirugia     otra_cirugia_txt  
##  Length:543         Min.   :0     Min.   :0.00000   Length:543        
##  Class :character   1st Qu.:0     1st Qu.:0.00000   Class :character  
##  Mode  :character   Median :0     Median :0.00000   Mode  :character  
##                     Mean   :0     Mean   :0.00208                     
##                     3rd Qu.:0     3rd Qu.:0.00000                     
##                     Max.   :0     Max.   :1.00000                     
##                     NA's   :63    NA's   :63                          
##   transfusion      transfusion_txt       unidades      expansores    
##  Min.   :0.00000   Length:543         Min.   :1.0   Min.   :0.0000  
##  1st Qu.:0.00000   Class :character   1st Qu.:1.0   1st Qu.:0.0000  
##  Median :0.00000   Mode  :character   Median :1.5   Median :0.0000  
##  Mean   :0.02083                      Mean   :1.5   Mean   :0.0375  
##  3rd Qu.:0.00000                      3rd Qu.:2.0   3rd Qu.:0.0000  
##  Max.   :1.00000                      Max.   :2.0   Max.   :1.0000  
##  NA's   :63                           NA's   :533   NA's   :63      
##  expansores_espec      otra            otra_txt          estado_mujer  
##  Min.   : 500     Min.   :0.00000   Length:543         Min.   :1.000  
##  1st Qu.:1000     1st Qu.:0.00000   Class :character   1st Qu.:1.000  
##  Median :1000     Median :0.00000   Mode  :character   Median :1.000  
##  Mean   :1941     Mean   :0.00417                      Mean   :1.023  
##  3rd Qu.:2500     3rd Qu.:0.00000                      3rd Qu.:1.000  
##  Max.   :4500     Max.   :1.00000                      Max.   :2.000  
##  NA's   :526      NA's   :63                           NA's   :63     
##  fecha_fallecio time_venoclisis               volumen_electrolitos
##  Min.   :NA     Min.   :2016-12-02 06:30:00   Min.   : 500        
##  1st Qu.:NA     1st Qu.:2017-02-16 11:38:30   1st Qu.:1000        
##  Median :NA     Median :2017-04-30 15:55:00   Median :1500        
##  Mean   :NA     Mean   :2017-04-29 16:02:30   Mean   :1479        
##  3rd Qu.:NA     3rd Qu.:2017-07-04 16:30:00   3rd Qu.:1925        
##  Max.   :NA     Max.   :2017-10-01 19:15:00   Max.   :6500        
##  NA's   :543    NA's   :67                    NA's   :67          
##     time_hb                      nivel_hb_4    totalbld_ARGoxy  totalbld500    
##  Min.   :2016-12-03 10:00:00   Min.   : 5.50   Min.   :  20    Min.   :0.0000  
##  1st Qu.:2017-02-16 10:07:30   1st Qu.: 9.60   1st Qu.: 150    1st Qu.:0.0000  
##  Median :2017-05-02 16:02:30   Median :10.50   Median : 300    Median :0.0000  
##  Mean   :2017-04-30 07:58:31   Mean   :10.38   Mean   : 385    Mean   :0.2222  
##  3rd Qu.:2017-07-04 15:51:45   3rd Qu.:11.20   3rd Qu.: 450    3rd Qu.:0.0000  
##  Max.   :2017-10-01 21:50:00   Max.   :15.40   Max.   :2000    Max.   :1.0000  
##  NA's   :67                    NA's   :67      NA's   :66      NA's   :66      
##   totalbld750      totalbld1000      totalbld1500       SI_baseline    
##  Min.   :0.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.4015  
##  1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.6375  
##  Median :0.000   Median :0.00000   Median :0.00000   Median :0.7297  
##  Mean   :0.109   Mean   :0.06709   Mean   :0.02725   Mean   :0.7422  
##  3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.8276  
##  Max.   :1.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.4568  
##  NA's   :66      NA's   :66        NA's   :66        NA's   :4       
##     SI_15min         SI_30min         SI_45min         SI_60min     
##  Min.   :0.4030   Min.   :0.3494   Min.   :0.3393   Min.   :0.3733  
##  1st Qu.:0.6943   1st Qu.:0.6704   1st Qu.:0.6569   1st Qu.:0.6569  
##  Median :0.8000   Median :0.7714   Median :0.7685   Median :0.7647  
##  Mean   :0.8253   Mean   :0.8138   Mean   :0.8083   Mean   :0.7987  
##  3rd Qu.:0.9144   3rd Qu.:0.8911   3rd Qu.:0.8934   3rd Qu.:0.8942  
##  Max.   :1.9138   Max.   :3.0000   Max.   :2.7778   Max.   :2.0000  
##  NA's   :63       NA's   :65       NA's   :66       NA's   :66      
##     gestage       prehbunder11       indoraug           tpd        
##  Min.   :25.86   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000  
##  1st Qu.:38.14   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 3.000  
##  Median :39.14   Median :0.0000   Median :0.0000   Median : 5.000  
##  Mean   :38.81   Mean   :0.2732   Mean   :0.2042   Mean   : 6.502  
##  3rd Qu.:40.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 7.000  
##  Max.   :42.14   Max.   :1.0000   Max.   :1.0000   Max.   :62.000  
##                  NA's   :5        NA's   :63       NA's   :65      
##      Hbdiff       HBdrop2ormore    Hbdrop2plusORbldtrans deliverytime     
##  Min.   :-6.100   Min.   :0.0000   Min.   :0.0000        Length:543       
##  1st Qu.: 0.400   1st Qu.:0.0000   1st Qu.:0.0000        Class1:hms       
##  Median : 1.200   Median :0.0000   Median :0.0000        Class2:difftime  
##  Mean   : 1.376   Mean   :0.3158   Mean   :0.3214        Mode  :numeric   
##  3rd Qu.: 2.300   3rd Qu.:1.0000   3rd Qu.:1.0000                         
##  Max.   : 6.800   Max.   :1.0000   Max.   :1.0000                         
##  NA's   :68       NA's   :68       NA's   :67                             
##  time_diagnosis time_prophylaxis  timebldcontrolled SI15min_threshold0.9
##  Min.   : 5.0   Min.   : 0.0000   Min.   : 0.0      Min.   :0.0000      
##  1st Qu.:15.0   1st Qu.: 0.0000   1st Qu.: 0.0      1st Qu.:0.0000      
##  Median :30.0   Median : 0.0000   Median :10.0      Median :0.0000      
##  Mean   :28.5   Mean   : 0.4217   Mean   :15.7      Mean   :0.2917      
##  3rd Qu.:33.0   3rd Qu.: 0.0000   3rd Qu.:25.0      3rd Qu.:1.0000      
##  Max.   :68.0   Max.   :26.0000   Max.   :72.0      Max.   :1.0000      
##  NA's   :438    NA's   :64        NA's   :438       NA's   :63          
##  SI30min_threshold0.9 SI45min_threshold0.9 SI60min_threshold0.9
##  Min.   :0.0000       Min.   :0.0000       Min.   :0.0000      
##  1st Qu.:0.0000       1st Qu.:0.0000       1st Qu.:0.0000      
##  Median :0.0000       Median :0.0000       Median :0.0000      
##  Mean   :0.2469       Mean   :0.2411       Mean   :0.2453      
##  3rd Qu.:0.0000       3rd Qu.:0.0000       3rd Qu.:0.0000      
##  Max.   :1.0000       Max.   :1.0000       Max.   :1.0000      
##  NA's   :65           NA's   :66           NA's   :66          
##  SI_delta_baseto15min SI_delta_baseto30min SI_delta_baseto45min
##  Min.   :-0.53695     Min.   :-0.51502     Min.   :-0.51950    
##  1st Qu.:-0.05481     1st Qu.:-0.07070     1st Qu.:-0.08544    
##  Median : 0.06632     Median : 0.03385     Median : 0.02392    
##  Mean   : 0.07891     Mean   : 0.06687     Mean   : 0.06178    
##  3rd Qu.: 0.18668     3rd Qu.: 0.15356     3rd Qu.: 0.15907    
##  Max.   : 1.23760     Max.   : 2.33962     Max.   : 2.06481    
##  NA's   :63           NA's   :65           NA's   :66          
##  SI_delta_baseto60min
##  Min.   :-0.55003    
##  1st Qu.:-0.08846    
##  Median : 0.02020    
##  Mean   : 0.05222    
##  3rd Qu.: 0.15467    
##  Max.   : 1.30189    
##  NA's   :66
```

Une hypothèse à tester

Plaçons-nous dans le contexte où l'on souhaite tester une hypothèse. Par exemple, supposons que l'on veuille tester si la moyenne de la variable `gestage`, l'âge gestationnel, est égale entre les mères qui avaient un historique d'hypertension et les autres (variable `hipertension_cronica`, égale à 1 si historique d'hypertension chronique). On pourrait par exemple vouloir déterminer si les mères souffrant d'hypertension chronique ont un âge gestationnel plus court, ou plus long.

On peut d'abord calculer la moyenne échantillonnale de l'âge gestationnel dans les deux groupes:

```
summary(data2[data2$hipertension_cronica==1,]$gestage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   36.43   37.39   38.36   38.36   39.32   40.29
```

```
summary(data2[data2$hipertension_cronica==0,]$gestage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   25.86   38.14   39.14   38.81   40.00   42.14
```

Les moyennes sont très similaires et la variance de l'âge gestationnel est relativement grande (ce résultat n'est pas montré), donc on s'attendrait ici à ce que le test ne rejette pas l'hypothèse nulle que les deux moyennes sont les mêmes. Voyons comment construire notre test plus en détails.

On doit d'abord poser les hypothèses nulle et alternative. Dans le cas mentionné ci-haut, on s'intéresse aux hypothèses suivantes:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

où μ_1 représente l'âge gestationnel moyen des mères sans historique d'hypertension et μ_2 celui des mères avec un historique d'hypertension. Notez que le test ci-haut est bilatéral puisqu'on ne spécifie pas dans l'hypothèse alternative si l'une des deux moyennes sera supérieure à l'autre.

La construction de tests d'hypothèse nous fournit une règle qui nous permet de faire un choix quant à l'hypothèse nulle, selon la statistique observée à partir de notre échantillon (ici, les deux moyennes échantillonnales seront utilisées dans le calcul de la statistique de test, qui consistera en une différence moyenne d'âge gestationnel standardisée).

On choisit ainsi une hypothèse nulle et on assume une distribution pour la statistique de test sous l'hypothèse nulle. C'est cette hypothèse nulle qu'on soumet à un test. Le test revient à évaluer la question suivante: *A quel point la statistique de test observée est peu probable sous la distribution engendrée par l'hypothèse nulle?* Notez que la statistique de test sous H_0 suit une distribution connue, que nous avons assumé (ex., une Student) qui dépend de paramètres. Souvent, le test revient à tester si ces paramètres sont égaux à certaines valeurs, mais on peut aussi remettre en doute la distribution assumée à prime abord. Ainsi, il faut généralement faire quelques vérifications (ex. histogramme pour voir la distribution des données) pour s'assurer que la distribution assumée pour la statistique de test elle-même a du sens.

Revenons à l'exemple de l'âge gestationnel. Soient $\bar{x}_1$ et $\bar{x}_2$ les moyennes échantillonnales d'âge gestationnel dans les groupes sans et avec historique d'hypertension, respectivement. Soient aussi s_1 et s_2 les écart-type échantillonnaux respectifs dans les deux groupes. La statistique de test pour une comparaison de moyennes revient à une différence moyenne standardisée (c'est intuitif: plus les moyennes échantillonnales s'éloignent entre elles, moins il est probable que $\mu_1 = \mu_2$). Voici la statistique pour ce test:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

où s_p est une variance pondérée calculée comme suit:

$$s_p = \sqrt{\frac{(n_2 - 1)s_2^2 + (n_1 - 1)s_1^2}{n_1 + n_2 - 2}}.$$

En assumant que les données $x_{1,1}, \ldots, x_{1,n_1}$ et $x_{2,1}, \ldots, x_{2,n_2}$ sont approximativement normales on peut montrer que la statistique T ci-haut suit une loi Student à $n_1 + n_2 - 2$ degrés de liberté. Ainsi, une statistique T observée qu'on considérerait "exagérée" ou trop loin de son centre serait une variable de loi Student à $n_1 + n_2 - 2$ degrés de liberté qui serait trop éloignée dans les ailes de la distribution.

Valeur-p

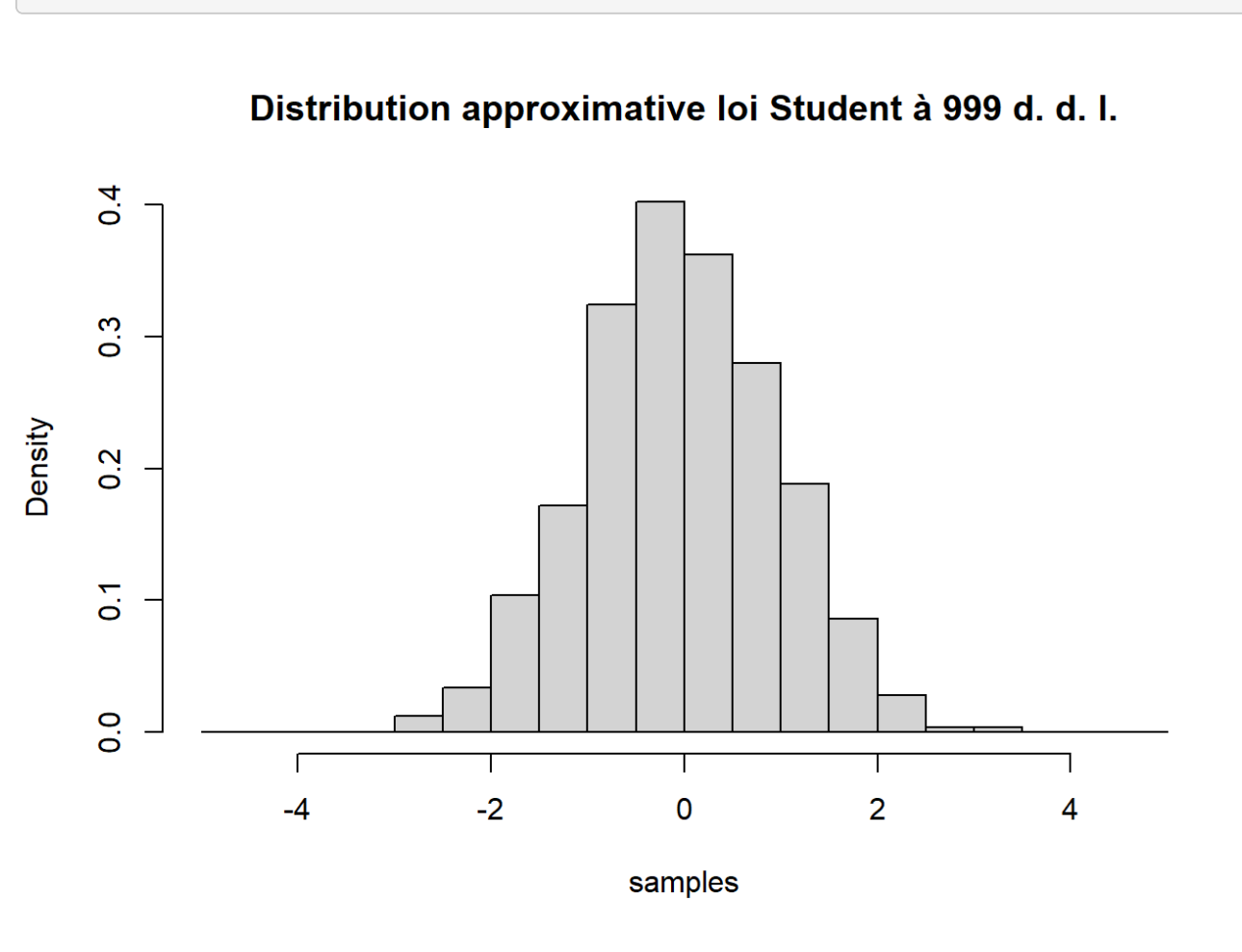
La valeur-p est intimement reliée à la statistique de test et au résultat du test. Elle représente la probabilité d'observer une statistique de test aussi ou plus "extrême" si l'hypothèse nulle est vraie. Ainsi, lorsque la valeur-p est petite, la statistique de test observée est peu probable sous H_0 et donc H_0 est peu probable (toujours en faisant l'hypothèse que la forme de la distribution, elle, est bien spécifiée).

Par exemple, pour la comparaison de deux moyennes, supposons qu'on ait finalement obtenu une statistique de test $T = 2$. Comme la statistique de test sous l'hypothèse nulle suit une distribution Student à $n_1 + n_2 - 2$ degrés de liberté (d. d. l.), on calculera la probabilité qu'une variable aléatoire de loi Student à $n_1 + n_2 - 2$ d. d. l. soit supérieure ou égale à 2, et on multipliera la réponse par 2 puisque la probabilité d'être inférieure ou égale à -2 doit être considérée comme aussi extrême pour cette statistique de test (c.-à-d. -2 est aussi loin du centre de la distribution que +2).

On peut le visualiser sur le graphique suivant:

```
n <- 1000 # 1000 unités
df <- n - 1 # degrés de liberté
samples <- rt(n, df)

hist(samples, breaks = c(-5,-4.5,-4,-3.5, -3,-2.5,-2,-1.5,-1,-0.5,0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5), freq = FALSE,
main='Distribution approximative loi Student à 999 d. d. l.')
```



Une statistique de test qui, sous H_0, suit une loi student similaire à celle ci-haut a une distribution concentrée autour de 0 et les chances de tomber sur une statistique de test observée supérieure à 2 ou inférieure à -2 sont relativement minces.

Minces à quel point? On comparera la valeur-p à un certain seuil, souvent 5%. Supposons par exemple qu'on obtienne une statistique de test $t = 2,4$ avec 99 degrés de liberté. Quelle est la valeur-p? Il s'agit de la probabilité pour une Student à 99 d. d. l. d'être supérieure à 2,4 ou inférieure à -2,4, qui se calcule comme ceci:

```
2*pt(q=-2.4, df=99,   lower.tail = TRUE, log.p = FALSE)
```

```
## [1] 0.01826457
```

pour une Student à 99 d. d. l. Notez que la fonction `pt` renvoie la probabilité d'être *inférieure* à la valeur mise dans `q` pour une Student à `df` d. d. l. Ici, nous multiplions donc par deux la probabilité d'être inférieur à -2,4, car cette probabilité est aussi égale à celle d'être supérieur à 2,4.

Puissance, niveau de significativité et erreurs de type I et II

Le test statistique est un outil qui nous permet de prendre une décision à partir de la statistique de test observée.

Encore une fois, la philosophie du test statistique est de postuler une hypothèse nulle puis de prouver (ou non) l'hypothèse alternative par l'absurde si on rejète l'hypothèse nulle. Le rejet de l'hypothèse nulle se fait après avoir observé une statistique de test trop extrême qui a peu de sens sous H_0.

Bien entendu, le test statistique n'est pas précis (ou correct) 100% du temps. On peut donc parler d'erreurs de type I ou II qui représentent respectivement les erreurs suivantes:

$$\alpha = \mathbb{P}(\text{Rejeter } H_0 | H_0 \text{ vraie})$$

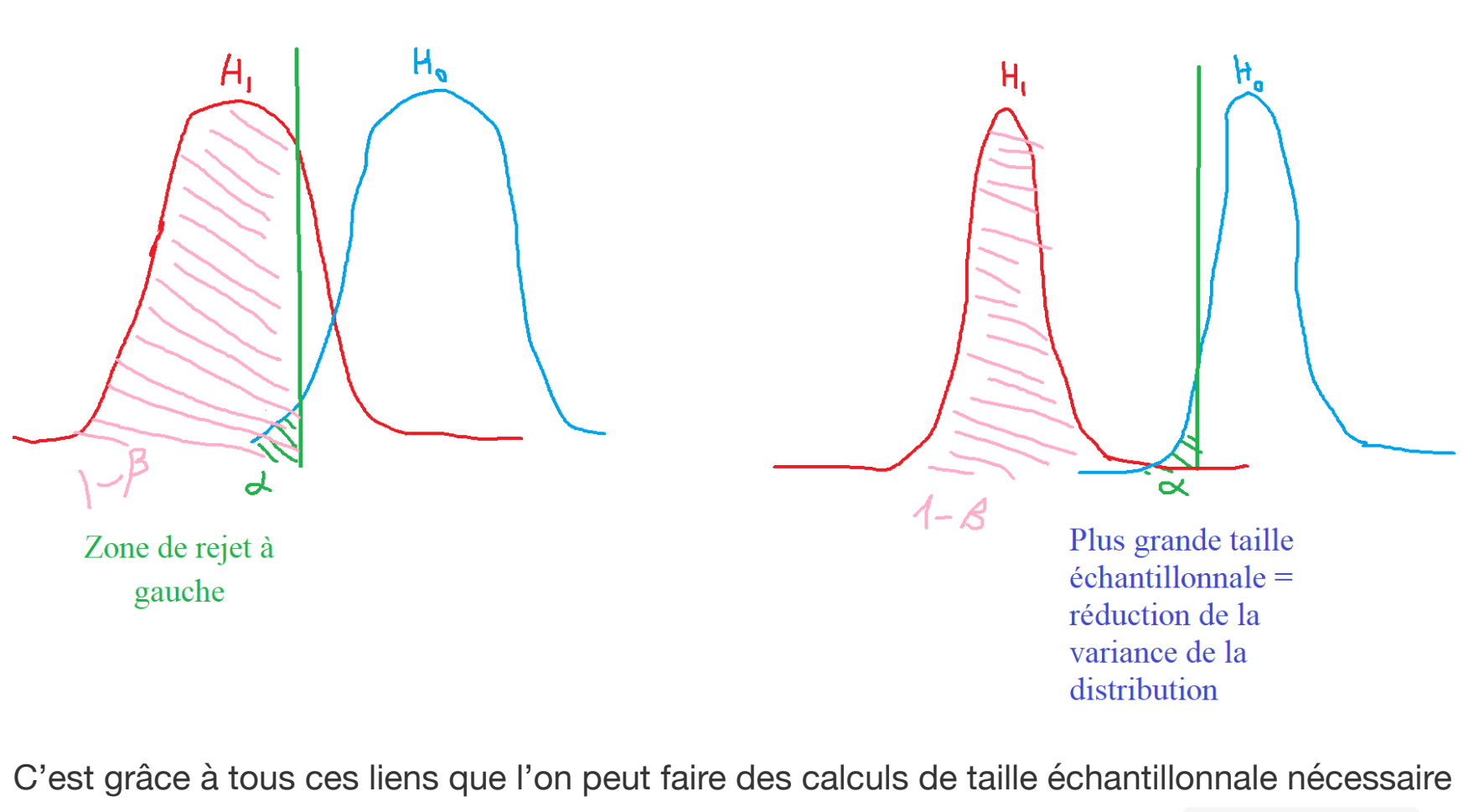
$$\beta = \mathbb{P}(\text{Ne pas rejeter } H_0 | H_0 \text{ fausse})$$

La puissance du test statistique est simplement donnée par $1 - \beta$ et représente donc la probabilité pour le test statistique de rejeter l'hypothèse nulle lorsqu'elle est effectivement fausse. Le niveau de significativité du test est égal à l'erreur de type I, qu'on a préalablement dénoté par α et il est fixé par le chercheur ou l'analyste (par nous). Dans la littérature, on retrouve souvent des valeurs de 5% et 1% pour α.

La puissance du test statistique, que l'on peut réécrire comme suit

$$\mathbb{P}(\text{Rejeter } H_0 | H_0 \text{ fausse}),$$

dépend de la taille échantillonale, puisque la taille de la zone de chevauchement où on a simultanément une réjection de H_0 (statistique de test observée trop extrême) et où H_1 est vraie (on est dans la distribution de H_1) dépend aussi de la taille échantillonnale (notez, on peut augmenter la puissance statistique en augmentant la taille d'échantillon - voir le graphique suivant).


H1
H0
1-β
α
Zone de rejet à gauche
H1
H0
1-β
α
Plus grande taille échantillonnale = réduction de la variance de la distribution

C'est grâce à tous ces liens que l'on peut faire des calculs de taille échantillonnale nécessaire pour obtenir une puissance statistique donnée. En R, des fonctions comme `power.t.test`, `pwr.2p.test` ou `pwr.f2.test` existent pour faire des calculs de puissance pour les tests-t, de proportion ou encore des modèles linéaires (je n'en discuterai pas plus en détails).

Notez que l'on n'a pas besoin de coder le test à la main ou de calculer la valeur-p à la main comme on a fait plus haut. Voyons maintenant en rafale quelques tests et comment les exécuter avec le logiciel R:

Tests pour variables continues

Test-t pour un échantillon

```
t.test( data2$gestage , mu=40 ) # Tester si l'âge gestationnel moyen est égal à 40
```

```
## 
##  One Sample t-test
## 
## data:  data2$gestage
## t = -14.18, df = 542, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 40
## 95 percent confidence interval:
##  38.64761 38.97670
## sample estimates:
## mean of x 
##  38.81215
```

Rejet de l'hypothèse nulle (valeur-p<0.05).

Test-t pour deux échantillons

```
hip<- data2[data2$hipertension_cronica==1,]
nonhip<-data2[data2$hipertension_cronica==0,]

t.test( hip$gestage, nonhip$gestage, var.equal=TRUE ) # Tester si les âges gestationnels moyens sont égaux sous l
'hypothèse de variance égale s1=s2
```

```
## 
##  Two Sample t-test
## 
## data:  hip$gestage and nonhip$gestage
## t = -0.33, df = 541, p-value = 0.7415
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.175221  2.261833
## sample estimates:
## mean of x mean of y 
##  38.35714  38.81384
```

Pas de rejet de l'hypothèse nulle (valeur-p>0.05).

Test-t pour deux échantillon avec correction de Welch (si l'on soupçonne que les variances ne sont pas similaires entre les groupes avec ou sans hypertension)

```
hip<- data2[data2$hipertension_cronica==1,]
nonhip<-data2[data2$hipertension_cronica==0,]

t.test( hip$gestage, nonhip$gestage, var.equal=FALSE ) # Tester si les âges gestationnels moyens sont égaux sous
l'hypothèse de variances inégales
```

```
##
##  Welch Two Sample t-test
##
## data:  hip$gestage and nonhip$gestage
## t = -0.23658, df = 1.0038, p-value = 0.852
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -24.76692  23.85354
## sample estimates:
## mean of x mean of y
##  38.35714  38.81384
```

Pas de rejet de l'hypothèse nulle (valeur-p>0.05).

Test pour vérifier si les variances sont égales (Test de Levene)

```
library(car)
leveneTest(data2$gestage ~ as.factor( data2$hipertension_cronica), data = data2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1   0.357 0.5504
##       541
```

Pas de rejet de l'hypothèse nulle (valeur-p>0.05). Notez que j'ai dû ajouter `as.factor()` autour de la variable groupe ici.

Test-t pour données appariées

On va d'abord créer de fausses données appariées, car on n'a pas de données appariées dans le jeu de données. Ainsi supposons que nous avons les mesures d'indice de masse corporelle (IMC) de 40 patients au temps 1 ainsi qu'au temps 2. Entre les deux temps, les patients ont reçu un nouveau programme d'exercice physique et on veut donc vérifier si les moyennes d'IMC ont changé.

```
temps1<- rnorm(n=40, mean= 32, sd=5)
temps2<- rnorm(n=40, mean=30, sd=10)
t.test( temps1, temps2, paired=TRUE ) # Tester si les IMC moyens sont égaux pour des données pairées ou corrélées
entre elles (ex: données au temps 1 vs au temps 2 pour une même personne)
```

```
##
##  Paired t-test
##
## data:  temps1 and temps2
## t = 1.2967, df = 39, p-value = 0.2023
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.194024  5.459530
## sample estimates:
## mean of the differences
##                2.132753
```

Pas de rejet de l'hypothèse nulle (valeur-p>0.05). Le test-t pour données appariées (ou pairées) tient compte de la corrélation entre les mesures au temps 1 et au temps 2 en testant si la différence moyenne (temps2 moins temps1) est différente de 0 ou non. En calculant cette différence et l'écart-type de cette différence dans l'échantillon, on n'a pas besoin de tenir compte de la covariance (temps1-temps2) de toute autre façon.

Tests pour variables binaires

Test-Z pour une proportion

On distingue deux façons de faire le test-Z en R, selon la taille échantillonnale. D'abord, si la taille échantillonnale est inférieure à 30, on utilise la fonction suivante:

```
x<- as.numeric(data2$cesarea)
full<- sum( x[is.na(x)==0]) ## garder ceux qui ne sont pas manquants
n<- length(x[is.na(x)==0]) ## total pour la binomiale
binom.test(full, n, p = 0.1, alternative= c("two.sided"))
```

```
##
##  Exact binomial test
##
## data:  full and n
## number of successes = 60, number of trials = 540, p-value = 0.389
## alternative hypothesis: true probability of success is not equal to 0.1
## 95 percent confidence interval:
##  0.08586744 0.14070090
## sample estimates:
## probability of success
##              0.1111111
```

Si, au contraire, la taille échantillonnale est supérieure à 30, alors on peut utiliser la fonction suivante:

```
x<- as.numeric(data2$cesarea)
full<- sum( x[is.na(x)==0]) ## garder ceux qui ne sont pas manquants
n<- length(x[is.na(x)==0]) ## total pour la binomiale
prop.test(full, n, p = 0.1, alternative= c("two.sided"), correct=TRUE) # pour correction de Yates pour continuité
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  full out of n, null probability 0.1
## X-squared = 0.62243, df = 1, p-value = 0.4301
## alternative hypothesis: true p is not equal to 0.1
## 95 percent confidence interval:
##  0.0864788 0.1414275
## sample estimates:
##         p
## 0.1111111
```

On différencie les cas de taille inférieure ou supérieur à 30, puisque l'un des deux tests (`prop.test`) nécessite de faire l'hypothèse de normalité de la moyenne (=proportion), qui est généralement à peu près vraie lorsque la taille échantillonnale passe au-dessus de 30. Le premier test (`binom.test`), quant à lui, ne nécessite pas de grande taille échantillonnale, puisqu'il est basé sur la distribution binomiale exacte et ne nécessite pas d'approximation normale.

Hypothèses associées avec le test-t

Les hypothèses associées sont les suivantes:

- Normalité des données dans l'échantillon (ou les deux échantillons si on fait le test-t pour deux échantillons)
- Pour le test à deux échantillons: Variances des groupes égales (on peut y remédier en utilisant le test pour variances inégales)
- Indépendance des observations (on peut y remédier si on a des paires, en utilisant le test-t pour donnée appariées)

On assume aussi, de base, que les données représentent des échantillons aléatoires provenant d'une plus grande population (ou de deux plus grandes populations) de sorte que la moyenne échantillonnale est sans biais pour la moyenne théorique.

L'hypothèse sur la normalité des données (par ex., pour deux échantillons) peut être évaluée à partir d'histogrammes. On souhaite retrouver des distributions unimodales et plutôt symétriques. Voyons voir pour la variable d'âge gestationnel:

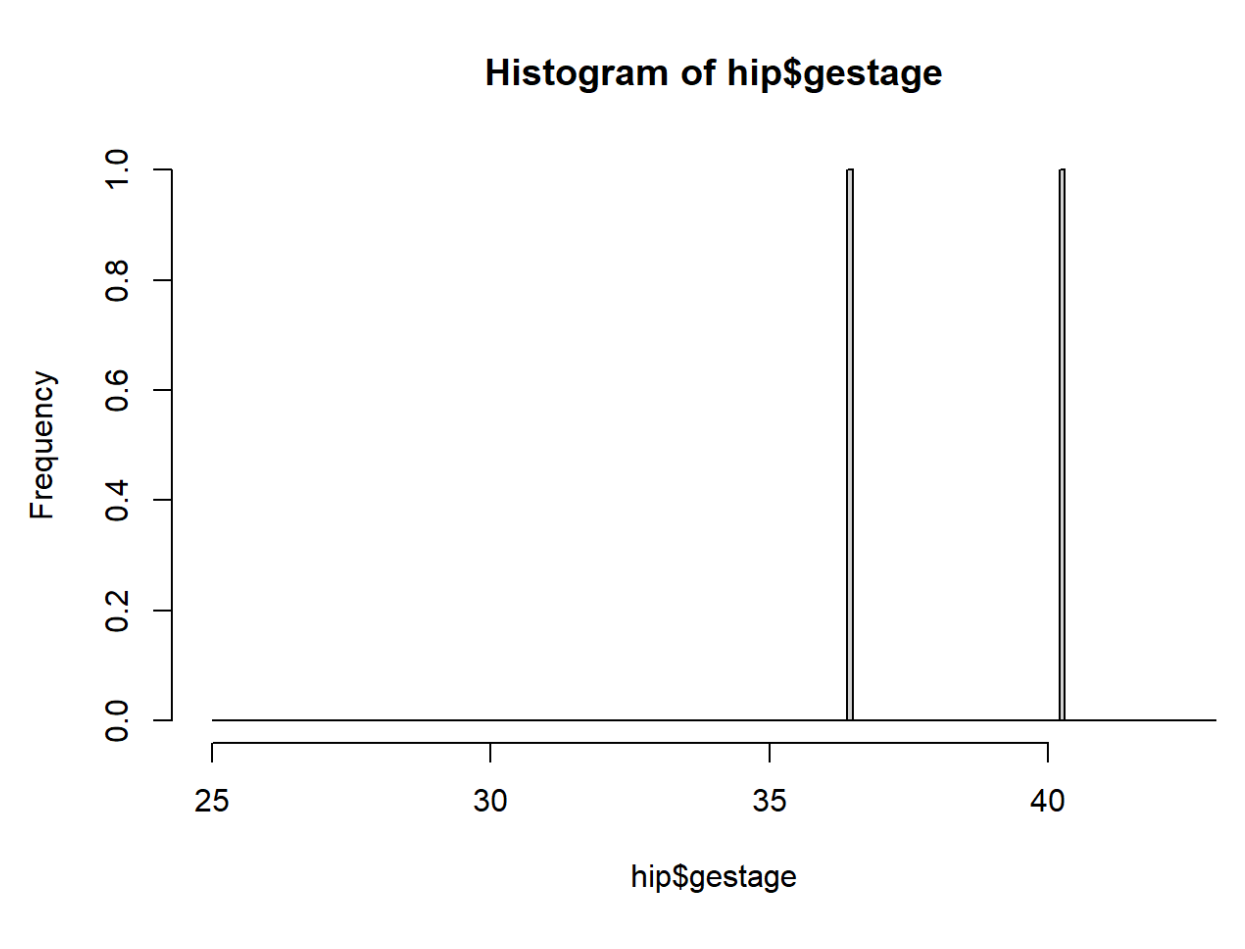
```
hip<- data2[data2$hipertension_cronica==1,]
nonhip<-data2[data2$hipertension_cronica==0,]
summary(hip$gestage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   36.43   37.39   38.36   38.36   39.32   40.29
```

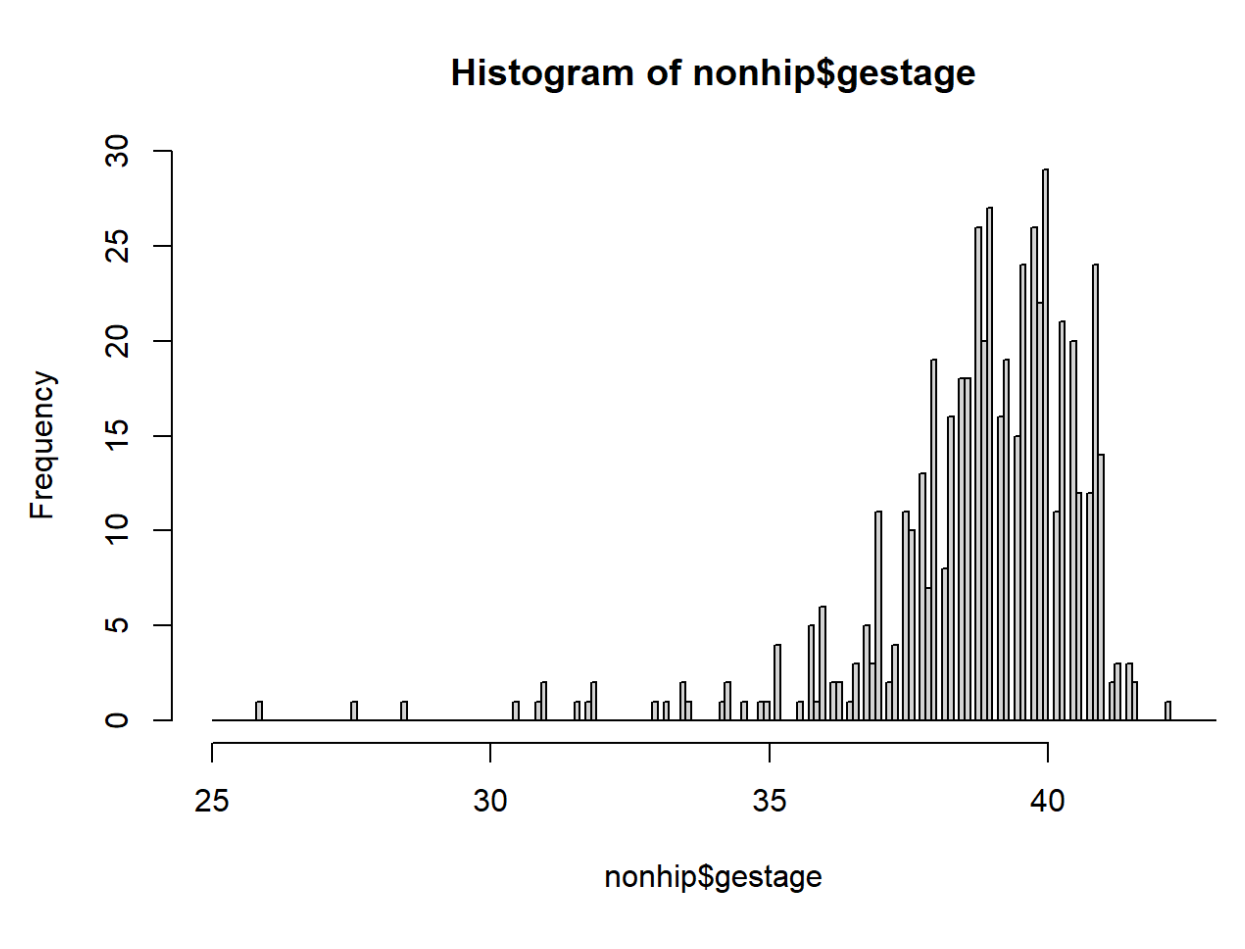
```
summary(nonhip$gestage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.86   38.14   39.14   38.81   40.00   42.14
```

```
hist(hip$gestage, breaks=c(seq(25,43,0.1)))
```



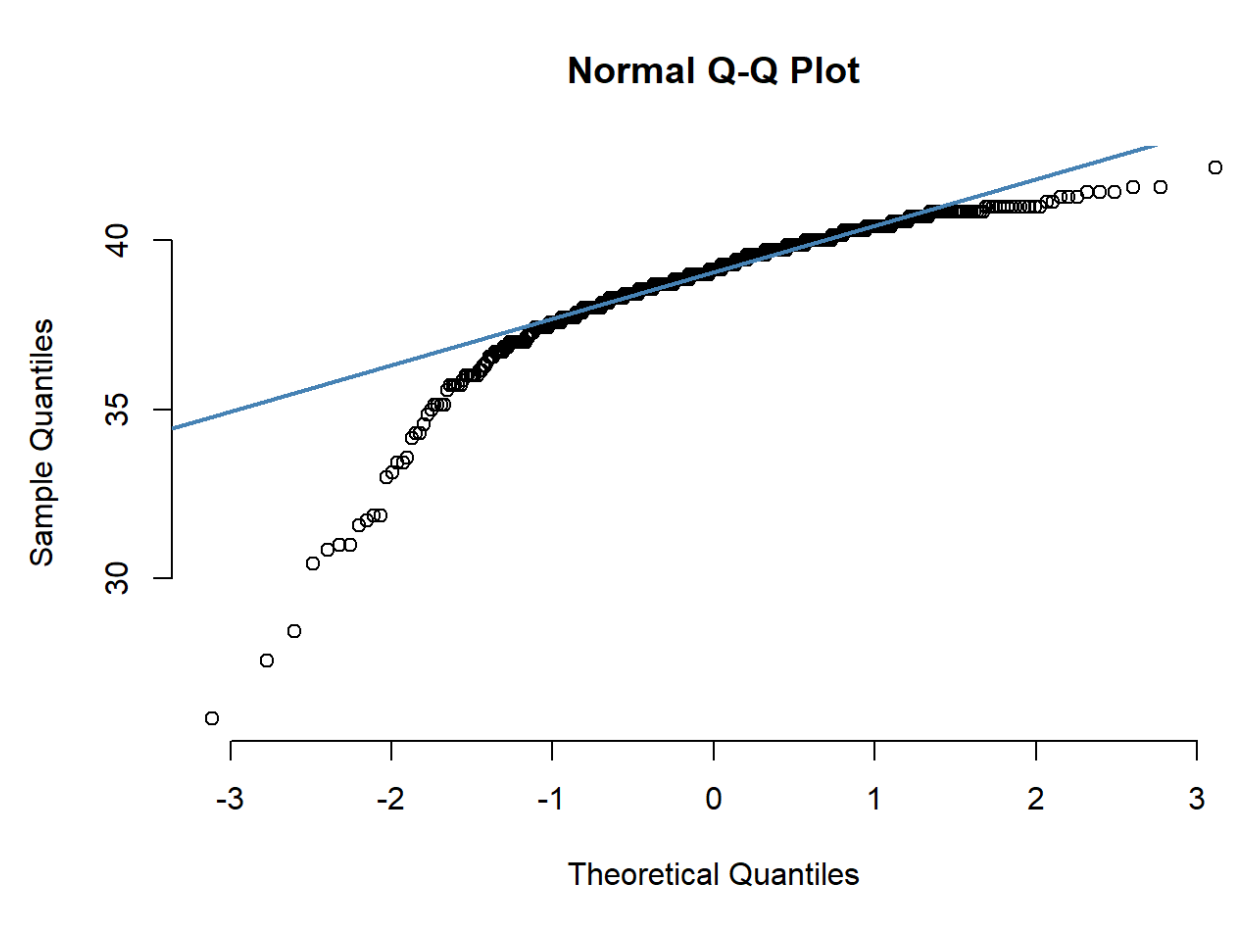
```
hist( nonhip$gestage, breaks=c(seq(25,43,0.1)))
```



Un des deux groupes présente avec une distribution de l'âge gestationnel plutôt Normale, qui est assez symétrique et plutôt unimodale (graphique du bas) alors que la distribution du haut est loin d'être distribuée comme une loi Normale. Dans ce cas, on doit se tourner vers des tests nonparamétriques qui n'assument pas la normalité des données (ex., le test de la somme des rangs de Wilcoxon).

Un autre diagnostic qu'on peut utiliser afin d'évaluer si l'hypothèse de normalité est raisonnable s'appelle le **QQplot**. Celui-ci nous montre la relation entre les quantiles théoriques de la loi Normale et les quantiles de la distribution de nos données. On peut le faire comme suit:

```
qqnorm(nonhip$gestage, pch = 1, frame = FALSE)
qqline(nonhip$gestage, col = "steelblue", lwd = 2)
```



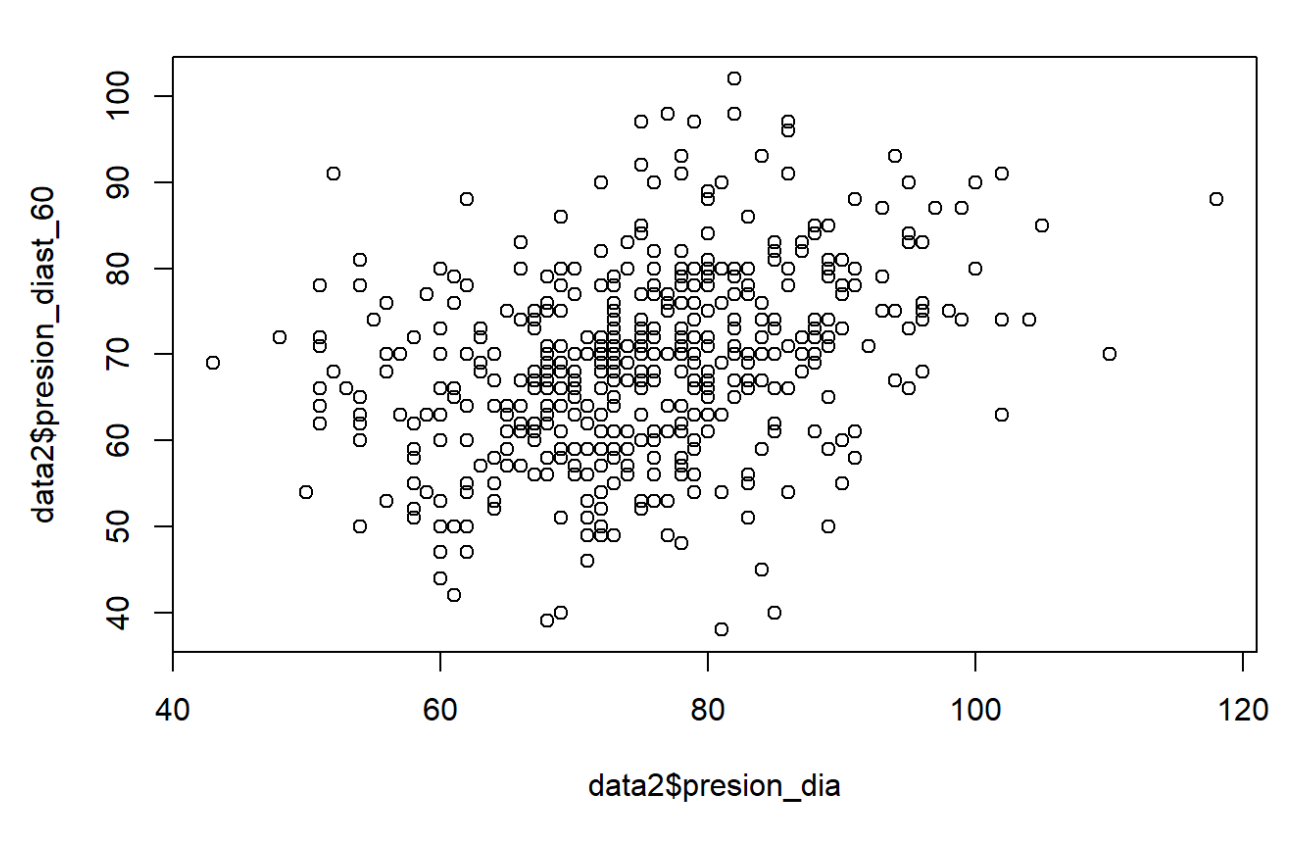
On souhaite voir des points près de la ligne droite, qui n'en dévient pas trop. Sur le QQplot pour le groupe sans hypertension ci-haut, on note de bonnes déviations. L'hypothèse de normalité n'est pas raisonnable dans ce cas.

Régression linéaire simple

La régression linéaire simple nous permet d'étudier une relation linéaire (association) entre deux variables continues. Par exemple, on peut s'intéresser au changement moyen dans le poids lorsqu'on change la consommation de calories de 100 unités par jour. Une régression linéaire qui modéliserait la relation entre le poids et le nombre de calories consommées nous permettrait de répondre à cette question.

Dans le jeu de données mentionné plus haut, supposons qu'on s'intéresse à la relation (qu'on assume linéaire) entre la pression diastolique pre partum (variable `presion_dia`) et la pression diastolique 60 minutes après l'accouchement (variable `presion_diast_60`). On peut décrire la relation grâce à un nuage de points:

```
plot( data2$presion_dia, data2$presion_diast_60  )
```



On note que plus la pression diastolique pre partum est élevée, plus la pression diastolique post partum l'est aussi (relation linéaire positive, c'est-à-dire, une pente droite ascendante).

Supposons maintenant que l'on s'intéresse à l'association entre une augmentation de la pression diastolique pre partum de 20 unités et l'augmentation moyenne de la pression pre partum. On peut obtenir cette statistique en ajustant une régression linéaire (fonction `lm` en R, pour `linear model`) et en spécifiant comme issue à gauche du modèle la variable que l'on assigne à l'axe des Y, et à droite du tilde (~), la variable que l'on assigne à l'axe des X:

```
lm(data2$presion_diast_60  ~ data2$presion_dia, data=data2)
```

```
##
## Call:
## lm(formula = data2$presion_diast_60 ~ data2$presion_dia, data = data2)
##
## Coefficients:
##       (Intercept)  data2$presion_dia
##           44.1398             0.3329
```

Ce que l'on note dans la sortie R, c'est une augmentation moyenne de 0,33 unité de la pression post partum lorsque l'on augmente la pression pre partum d'une unité. Ainsi, pour 20 unités, on peut multiplier par 20, ce qui nous donne

```
20*0.3329
```

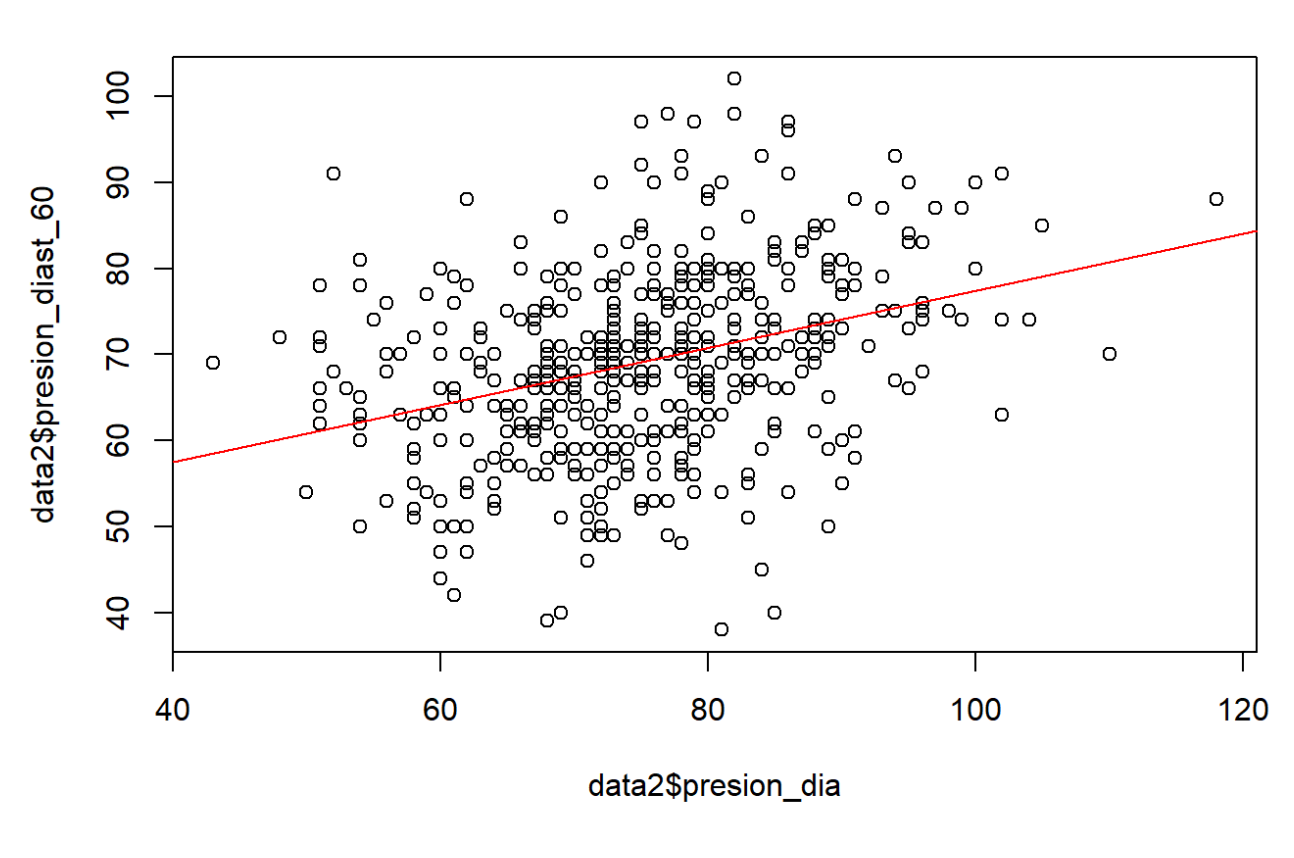
```
## [1] 6.658
```

une augmentation moyenne de 6.7 unités dans la pression diastolique post partum pour chaque saut de 20 unités dans la pression pre partum. Bien sûr, il ne s'agit que d'une association et non pas un lien de cause à effet, qui nécessiterait d'autres hypothèses.

La sortie plus haut peut être interprétée comme suit: On peut estimer la pression post partum moyenne (ou espérée) en calculant `44.1398 + 0.3329*pression pre partum`. On obtient, à partir de cette formule, une approximation de la pression post partum basée sur la droite de régression ajustée et le prédicteur de pression pre partum.

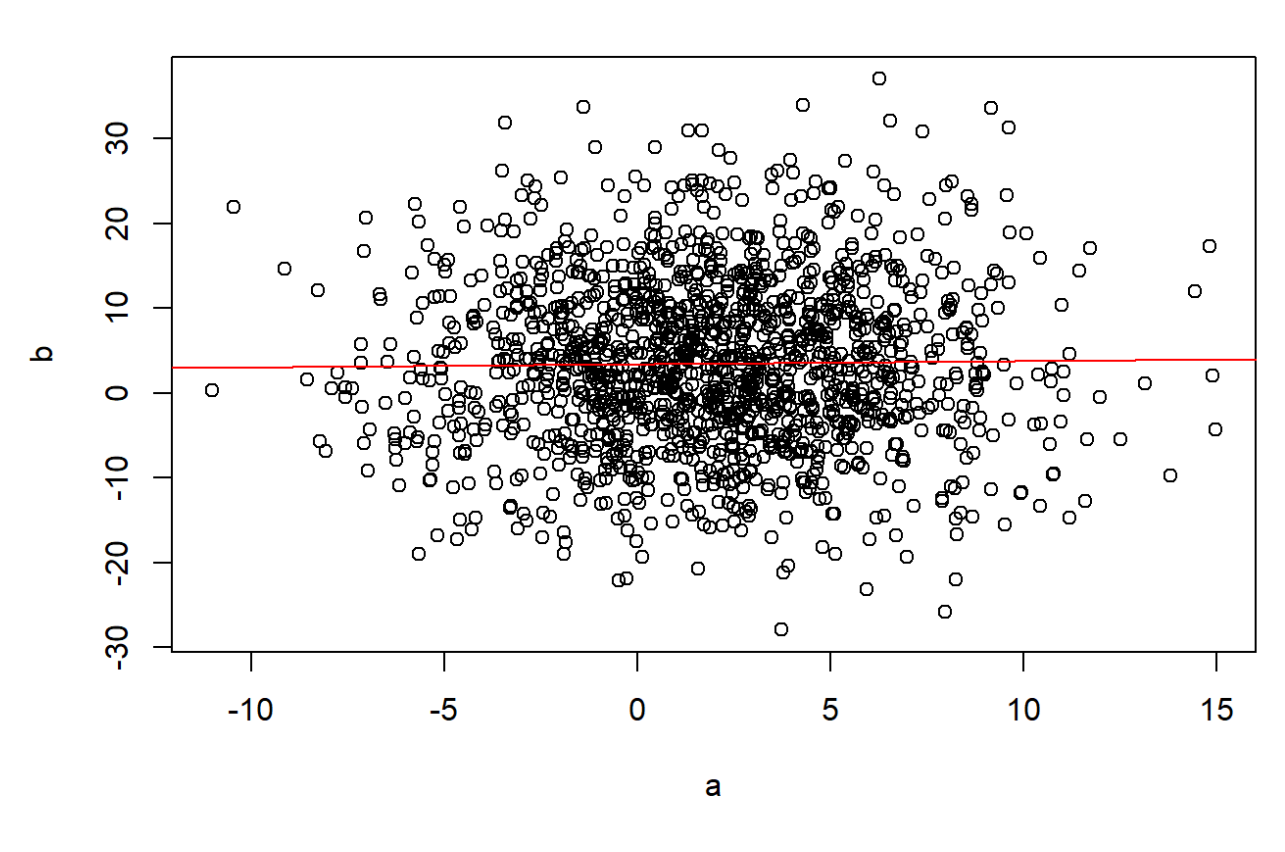
Visualisons cette droite de régression par-dessus le nuage de points:

```
plot( data2$presion_dia, data2$presion_diast_60  )
abline(lm(data2$presion_diast_60 ~ data2$presion_dia), col='red')
```



Lorsque deux variables ne sont pas associées, on peut s'attendre à voir un nuage de points ne présentant pas de tendance ascendante ou descendante, c'est-à-dire, quelque chose comme ça (de plat):

```
a<- rnorm( n=1500, mean= 2, sd=4)
b<- rnorm( n=1500, mean= 3.5, sd=10)
plot( a , b  )
abline(lm(b  ~ a), col='red')
```

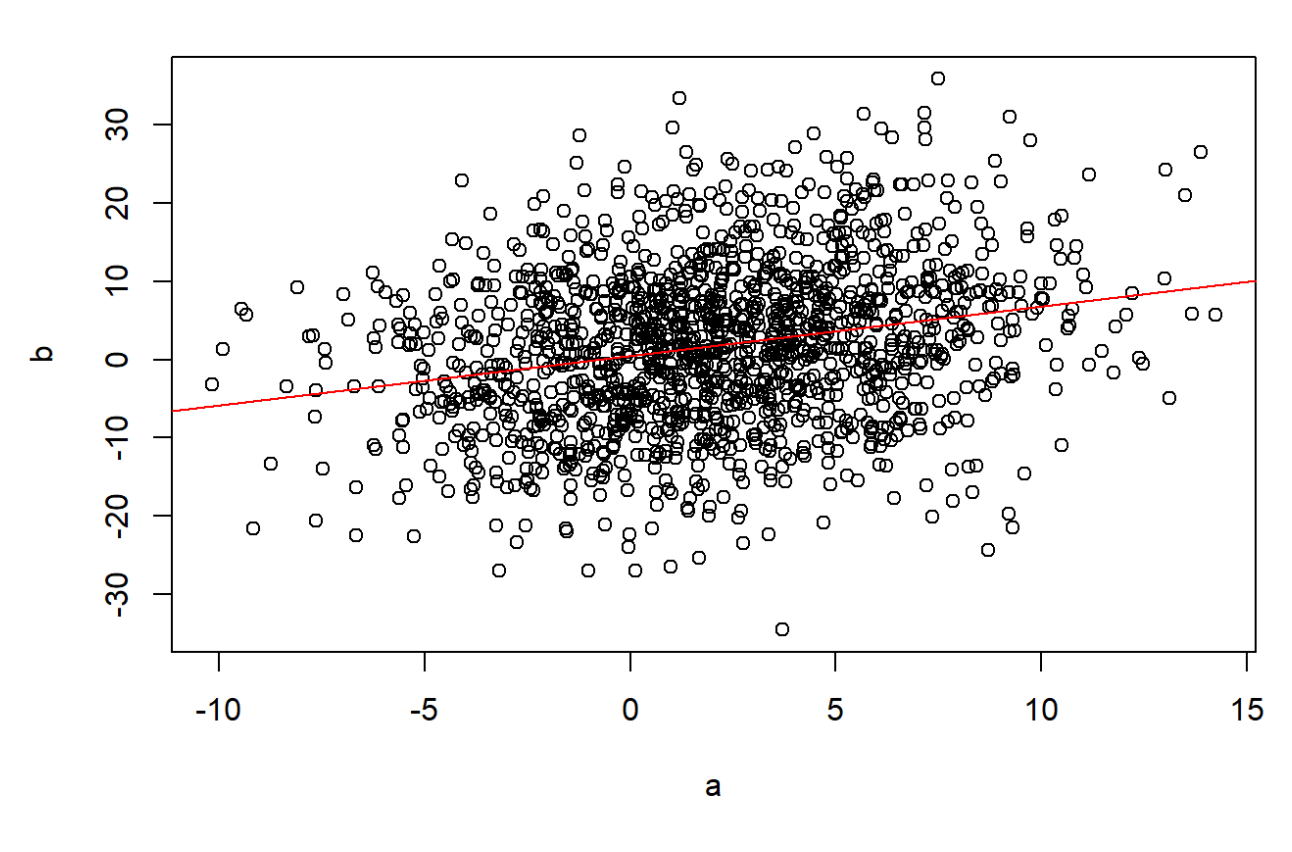


L'intuition derrière ce nuage de points plat? On peut dire que peu importe la valeur de la variable sur l'axe des X, on ne peut pas vraiment inférer sur la valeur de la variable sur l'axe des Y, et vice-versa (il n'y a pas d'association).

Lorsque deux variables sont associées, elles peuvent être associées **de façon linéaire**, ou encore, d'une autre façon (ex.: quadratique). Voici des exemples de deux variables qui sont associées linéairement:

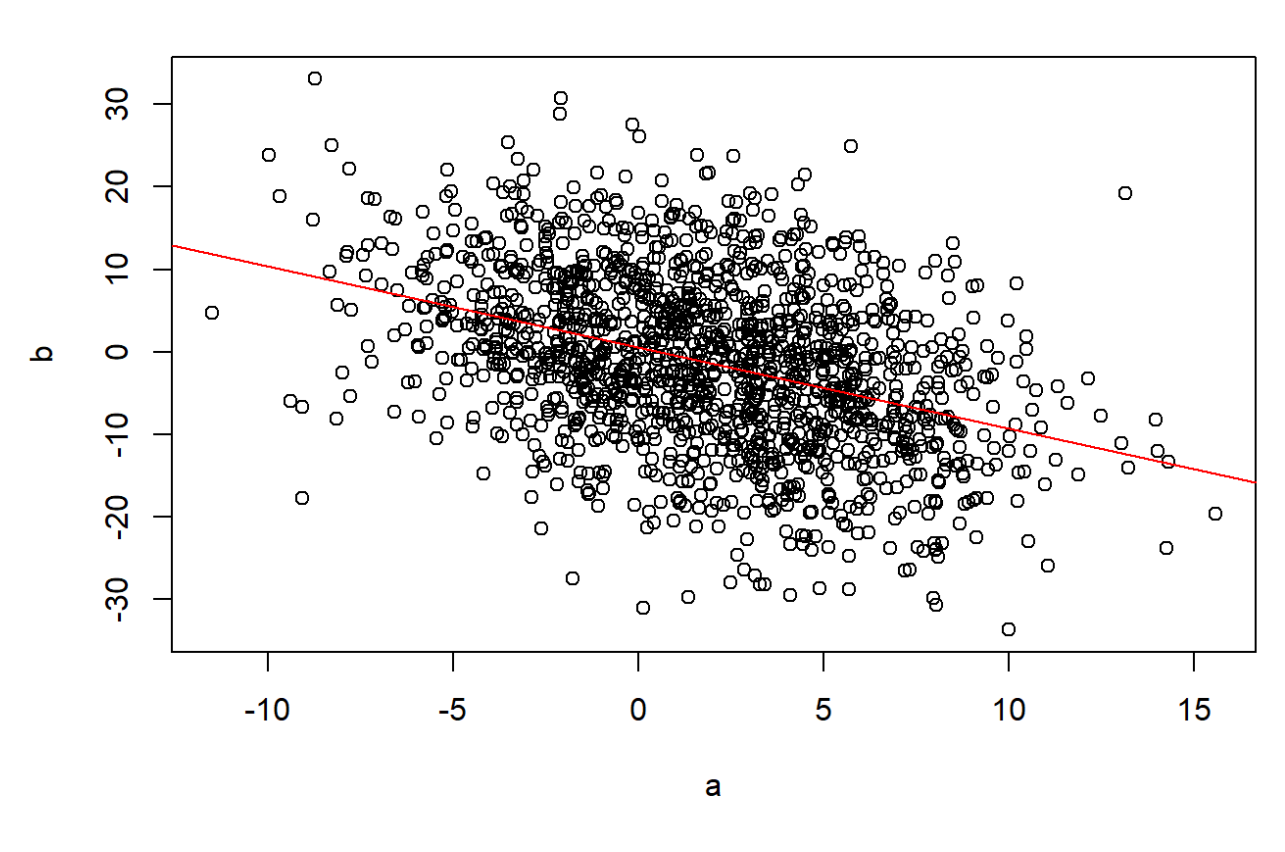
Association linéaire positive

```
a<- rnorm( n=1500, mean= 2, sd=4)
b<- rnorm( n=1500, mean= 0.7*a, sd=10)
plot( a , b  )
abline(lm(b  ~ a), col='red')
```



Association linéaire négative

```
a<- rnorm( n=1500, mean= 2, sd=4)
b<- rnorm( n=1500, mean= -0.9*a, sd=10)
plot( a , b  )
abline(lm(b  ~ a), col='red')
```



On représente typiquement la droite de régression linéaire à partir d'une ordonnée à l'origine (β_0) et d'une pente (β_1). Chaque point sur la droite de régression, vis-à-vis un x donné, est égal à

$$y = \beta_0 + \beta_1 x.$$

Vérifier s'il existe une association linéaire entre deux variables revient à tester si la pente de régression linéaire est non-nulle (puisqu'une pente nulle, donc égale à 0, mène à une droite de régression horizontale constante qui est égale à l'ordonnée à l'origine β_0 partout).

Ainsi, le test statistique permettant de vérifier si deux variables sont associées de façon linéaire est le suivant:

$$H_0 : \beta_1 = 0.$$

En R, pour faire le test, il suffit d'ajuster la droite de régression linéaire et de demander le résumé de l'ajustement, comme suit:

```
ma_regression<- lm(data2$presion_diast_60 ~ data2$presion_dia)
summary(ma_regression)
```

```
##
## Call:
## lm(formula = data2$presion_diast_60 ~ data2$presion_dia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.102  -6.442  -0.095   7.220  30.565
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       44.13976    3.20062  13.791   <2e-16 ***
## data2$presion_dia  0.33287    0.04215   7.897    2e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.28 on 475 degrees of freedom
##   (66 observations deleted due to missingness)
## Multiple R-squared:  0.1161, Adjusted R-squared:  0.1142
## F-statistic: 62.36 on 1 and 475 DF,  p-value: 1.997e-14
```

La valeur-p du test de pente nulle est égale à 2e-14, dans ce cas-ci (c'est la valeur-p vis-à-vis la variable `data2$presion_dia`). Ainsi, on rejette l'hypothèse nulle qu'il n'y a pas d'association linéaire entre les pressions pre et post partum dans cette étude.

Hypothèses associées avec la régression linéaire simple

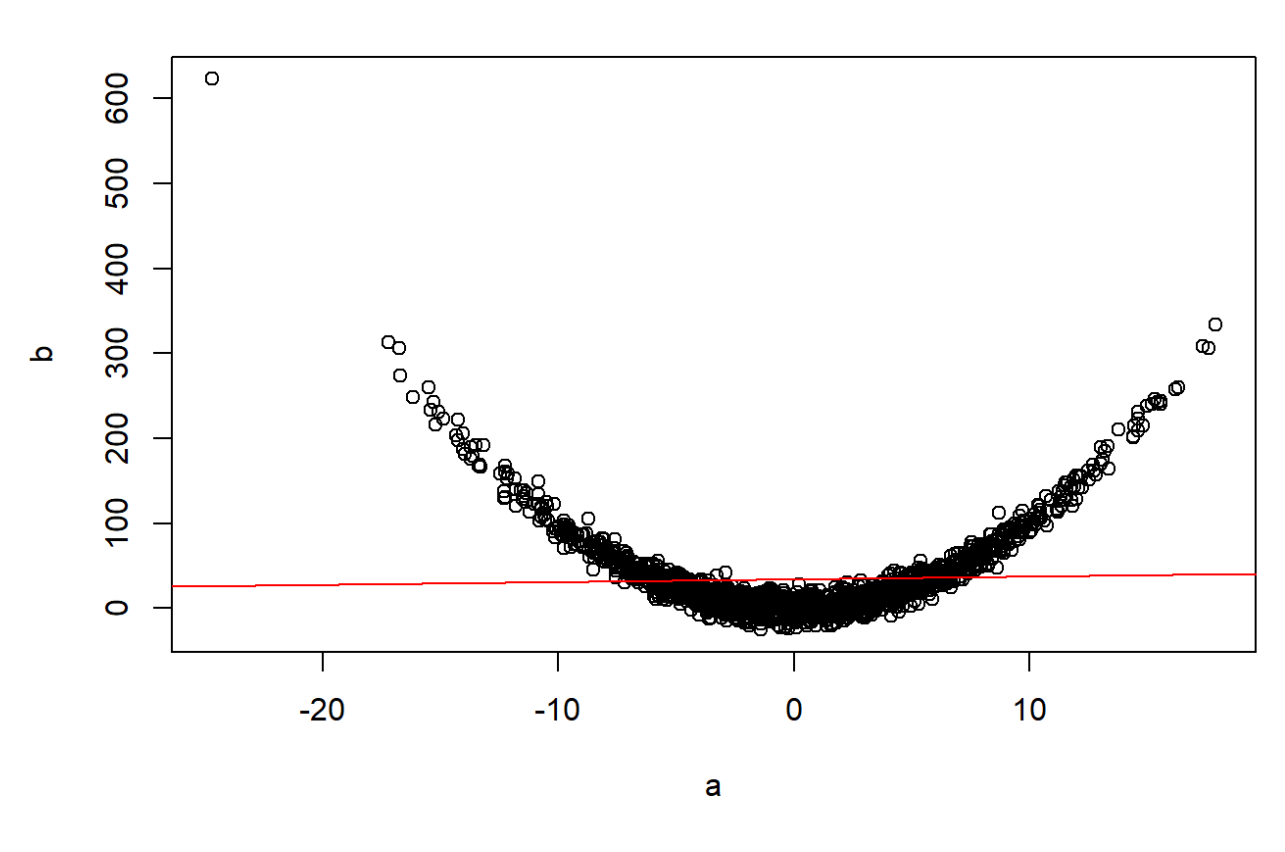
La régression linéaire simple se base sur les hypothèses suivantes:

- Il y a une relation linéaire entre les deux variables
- Normalité des données (qui nous permet d'avoir une statistique de test, de connaître la loi de cette statistique de test, de pouvoir faire un test statistique, puis d'obtenir une valeur-p)
- Données indépendantes entre elles

Notez que la régression linéaire simple discutée ici peut être étendue au cas où on a plus d'un prédicteur (ce qu'on appelle la régression linéaire multiple). Dans ce cas, on doit assumer que les différents prédicteurs ne sont pas multicolinéaires, c'est-à-dire qu'ils ne sont pas associés linéairement entre eux.

Note. Qu'arrive-t-il si la relation entre les variables n'est pas linéaire? Il se peut qu'il existe une relation très forte entre deux variables, mais que cette relation soit plutôt quadratique (voir par exemple le graphique suivant)

```
a<- rnorm( n=1500, mean= 0, sd=6)
b<- a^2 + rnorm( n=1500, mean= 0, sd=10)
plot( a , b  )
abline(lm(b  ~ a), col='red')
```



Dans ce cas, notez que l'on ne rejettera fort probablement pas l'hypothèse nulle que $\beta_1 = 0$ puisque la droite de régression linéaire semble avoir une pente d'environ 0. Par contre, on a une relation forte entre les deux variables. Ainsi, le fait de ne pas rejeter l'hypothèse nulle qu'il n'y a pas de relation linéaire **ne signifie pas** qu'il n'y a pas d'autre relation entre les variables.

Finalement, notez que le même type de diagnostique que pour le test-t (ex., QQplot et histogramme) peut être utilisé pour juger de la normalité des données, une des hypothèses de la régression linéaire.

Exercice

Pour cet exercice, on utilisera le jeu de données `ChickWeight` fourni en R, qu'on peut télécharger sur R grâce à la commande suivante:

```
data(ChickWeight)
```

Ce jeu de données contient 4 variables en lien avec le poids, le temps, l'identifiant du poussin et sa diète. Je montre ici les 10 premières lignes du jeu de données ainsi qu'un résumé des 4 variables:

```
head(ChickWeight, 10)
```

```
##    weight Time Chick Diet
## 1      42    0     1    1
## 2      51    2     1    1
## 3      59    4     1    1
## 4      64    6     1    1
## 5      76    8     1    1
## 6      93   10     1    1
## 7     106   12     1    1
## 8     125   14     1    1
## 9     149   16     1    1
## 10    171   18     1    1
```

```
summary(ChickWeight)
```

```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
##  Max.   :373.0   Max.   :21.00   19     : 12
##                                  (Other):506
```

1. Choisissez un test qui vous permettra de comparer le poids moyen des poussins sous la diète #1 et la diète #2. Conduisez le test et rapportez les résultats et votre conclusion concernant la comparaison des deux diètes.
2. Choisissez un diagnostique pour évaluer la normalité des données utilisées en question 1. Faites le diagnostique et discutez des résultats. Est-ce que l'hypothèse de normalité semble tenir ici?
3. On s'intéresse à la relation entre le poids et le temps. Supposons que le temps et le poids sont toutes deux des variables continues et ajustons une régression linéaire pour étudier le poids (axe des Y) en relation au temps (axe des X). Utilisez la fonction `plot` pour voir la droite de régression linéaire simple par-dessus le nuage de points. À partir du résultat provenant de `lm` et du graphique, que concluez-vous? (note. Il est normal que le nuage de points formera des bandes verticales, puisque la variable temps du jeu de données n'est pas vraiment continue ici).

[1] Durocher J. A double-blind, randomized controlled trial on oxytocin route for prevention of postpartum hemorrhage in Argentina 2019. https://doi.org/10.7910/DVN/MDZRKU.