Que faire avec les variables semi-quantitatives ? Quelques possibilités abordées à l'aide d'exemples.

Marie Laure Delignette-Muller

14 février, 2022

Que faire avec des données semi-quantitatives ?

Abordons quelques possibilités à l'aide de données "jouet" issues d'une enquête réalisée par les étudiants de S6 en mai 2017 auprès de leur propre promotion.

```
d <- read.table("DATA/enquete1617.txt", header = TRUE,
                stringsAsFactors = TRUE)
str(d)
```

```
## 'data.frame':    93 obs. of  27 variables:
##  $ Q1sexe          : Factor w/ 2 levels "femme","homme":
##  $ Q2etudes        : Factor w/ 2 levels "non","oui": 1 2
##  $ Q3sante         : Factor w/ 2 levels "non","oui": 1 2
##  $ Q4agricole      : Factor w/ 2 levels "non","oui": 1 1
##  $ Q5liberal       : Factor w/ 2 levels "non","oui": 2 2
##  $ Q6tempstravail  : int  50 50 43 45 40 45 42 36 50 39
##  $ Q7choixliberal  : Factor w/ 2 levels "liberal","salari
##  $ Q8bourse        : Factor w/ 2 levels "non","oui": 2 2
##  $ Q9job           : num  0 0 4 0 0 0 0 4 0 4 ...
##  $ Q10animal       : Factor w/ 2 levels "non","oui": 2 2
```

Variables sur lesquelles nous allons travailler ?

```
str(d[, c(8, 9, 10, 12, 21, 22, 24, 26)])
```

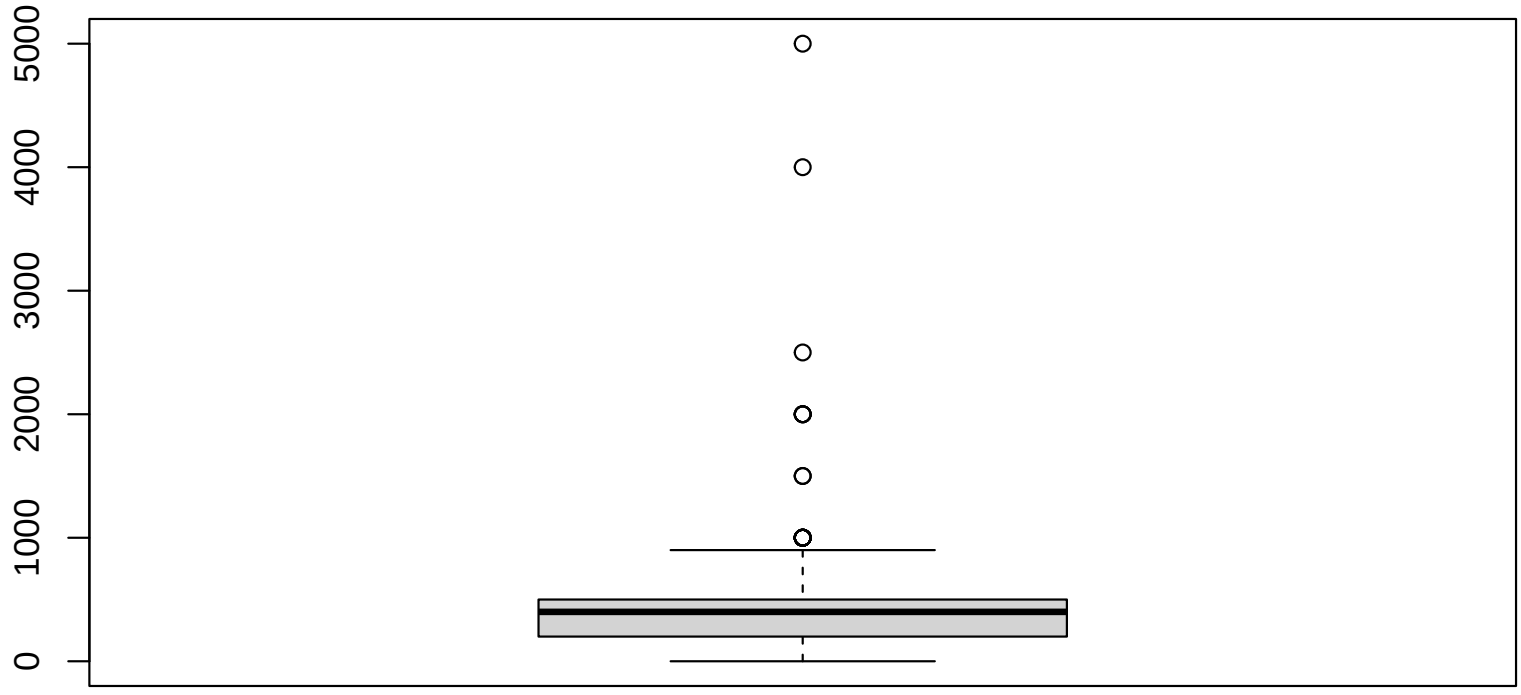
```
## 'data.frame':    93 obs. of  8 variables:
##  $ Q8bourse    : Factor w/ 2 levels "non","oui": 2 2 2 1
##  $ Q9job       : num  0 0 4 0 0 0 0 4 0 4 ...
##  $ Q10animal   : Factor w/ 2 levels "non","oui": 2 2 1 1
##  $ Q12coutsoins: int  1000 300 1000 300 100 500 1000 100
##  $ Q21clubs    : int  7 5 4 3 5 2 2 7 2 7 ...
##  $ Q22amphi    : int  2 1 0 1 2 3 1 7 1 8 ...
##  $ Q24km       : num  800 450 400 650 480 ...
##  $ Q26polys    : Factor w/ 3 levels "milieu","tard",..:
```

Cas d'une variable quasi quantitative continue

Distribution du montant maximal acceptable pour des soins vétérinaires

Q12 : Quel montant maximal (en euros) vous paraît-il acceptable de débourser pour soigner un chat ou un chien ?

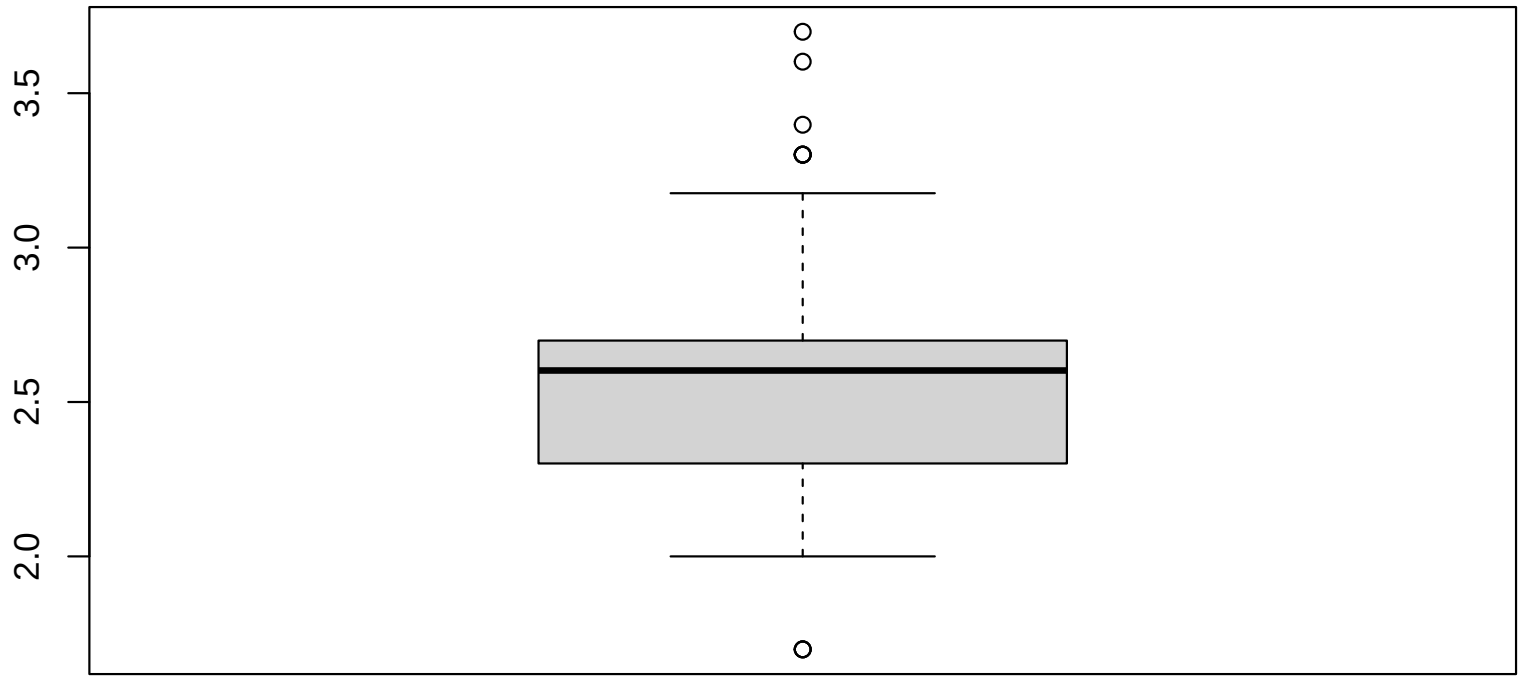
```
boxplot(d$Q12coutsoins)
```



Essai de transformation logarithmique

Peut-on tenter une transformation logarithmique pour normaliser la distribution ?

```
boxplot(log10(d$Q12coutsoins))
```



Quel est le problème ?

Regardons de plus près les données pour comprendre le “warning”

```
summary(d$Q12coutsoins)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     200     400     615     500    5000
```

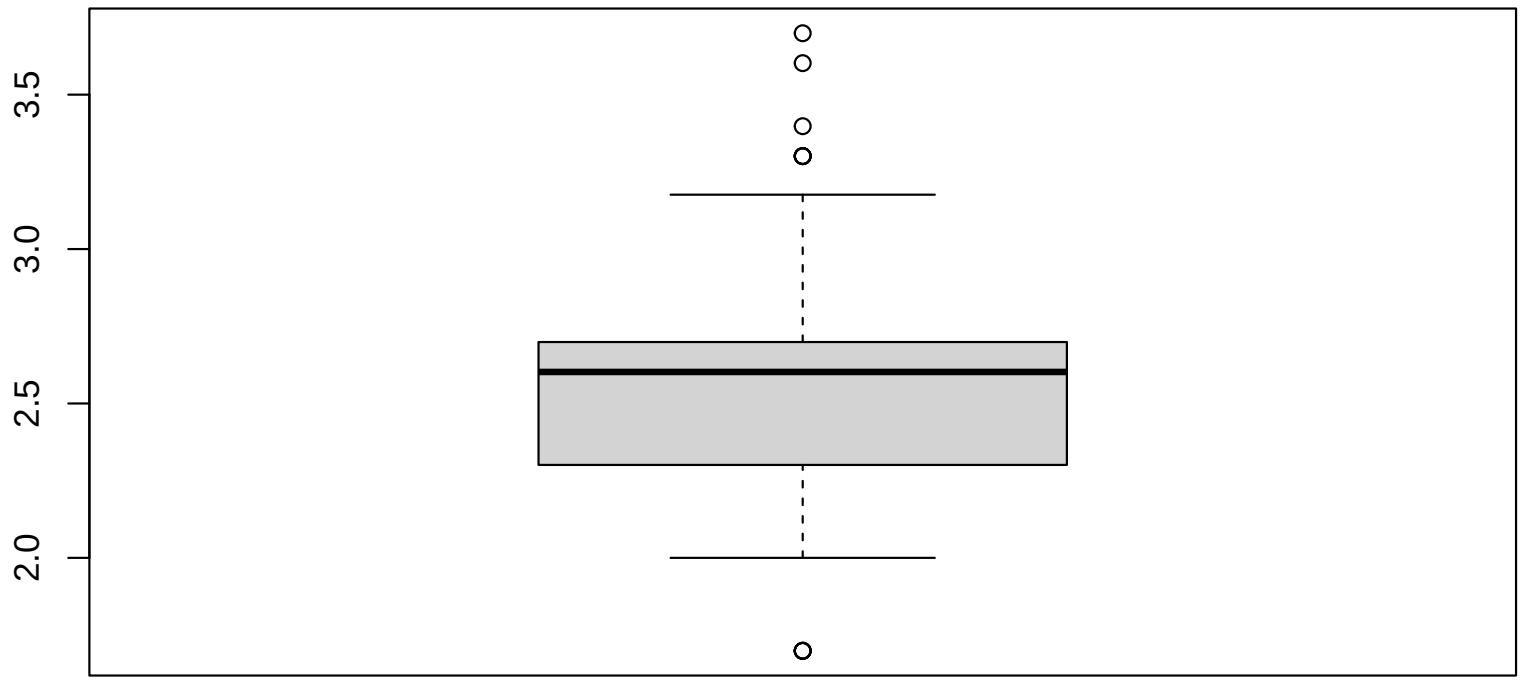
Certains ont répondu 0. Cela semble un peu “gonflé” comme réponse pour des futurs vétérinaires. Par quelle valeur pourrrait-on remplacer les 0 (qui ne sont sans doute pas des vrais 0) en tenant compte qu'ils sont dans la queue de distribution à gauche ?

Imaginons que ce sont des valeurs censurées à gauche et remplaçons-les par la moitiée de plus petite valeur non nulle donnée.

```
minnonnul <- min(d$Q12coutsoins[d$Q12coutsoins != 0])
d$Q12coutsoins[d$Q12coutsoins == 0] <- minnonnul
```

Transformation logarithmique (2)

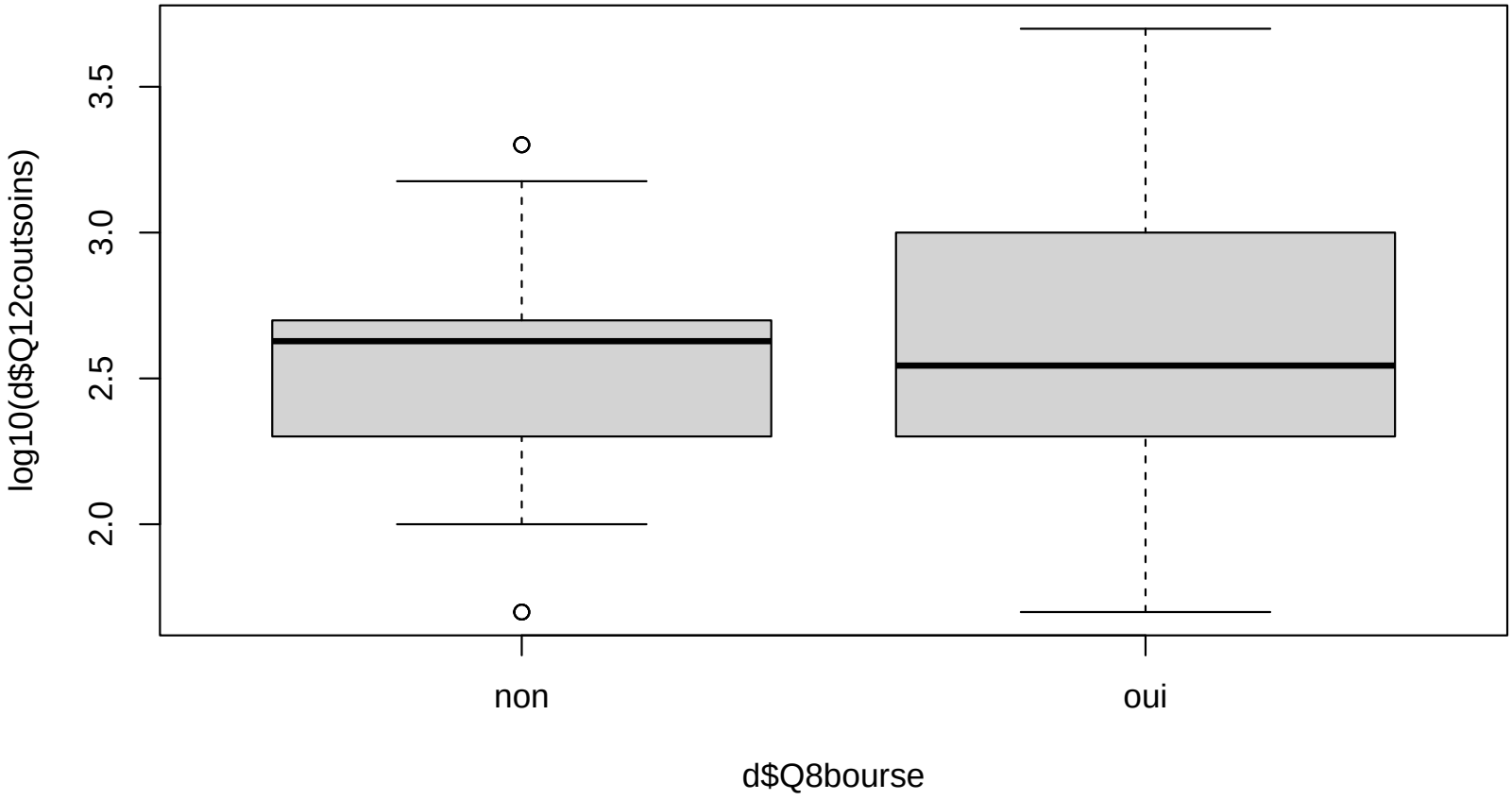
```
boxplot(log10(d$Q12coutsoins))
```



Corrélation de la question Q12 à la question Q8

Q8 : Bénéficiez-vous d'une bourse sur critères sociaux ?

```
plot(log10(d$Q12coutsoins) ~ d$Q8bourse)
```



Visualisation de l'ensemble des points

```
stripchart(log10(d$Q12coutsoins) ~ d$Q8bourse,
           vertical = TRUE, method = "stack")
```

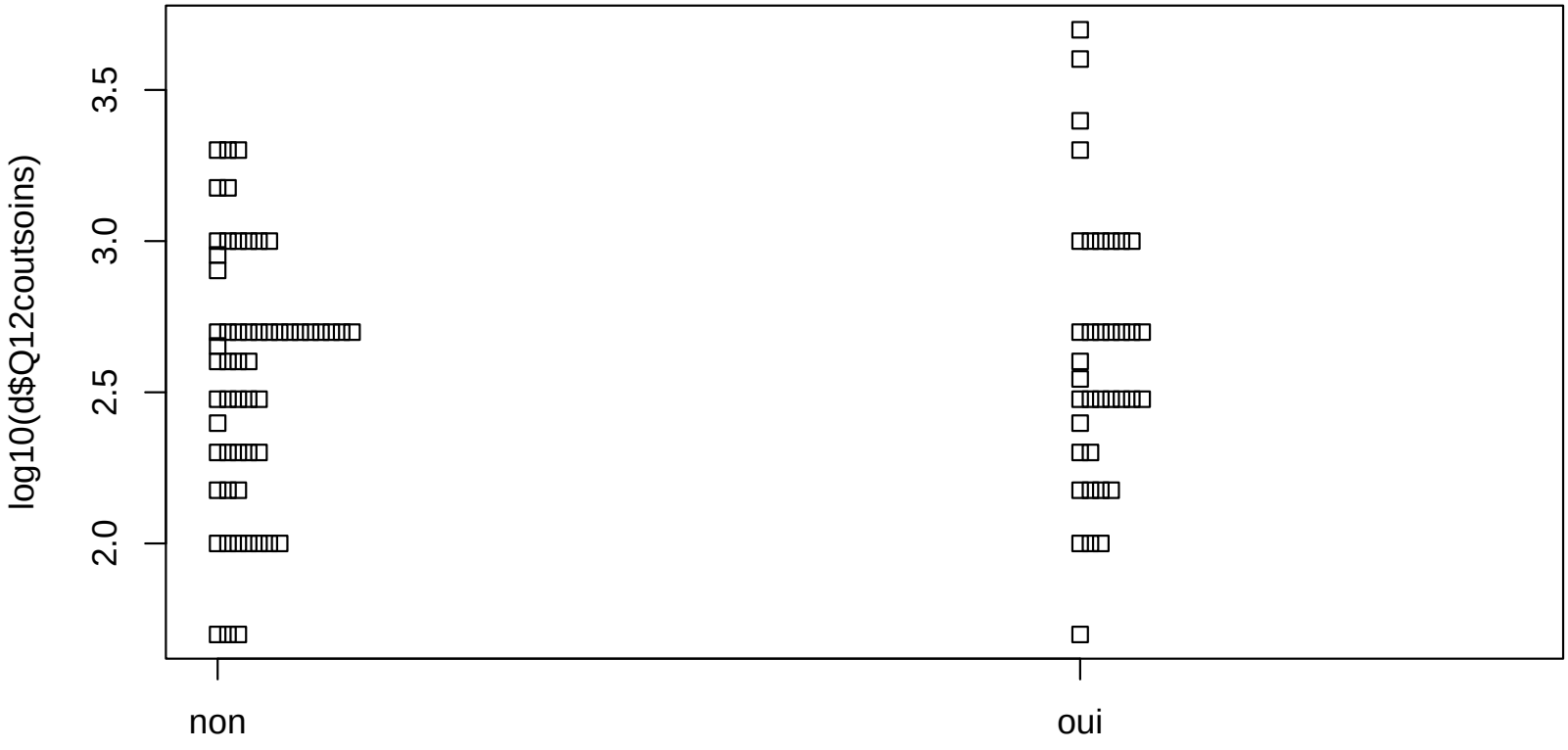
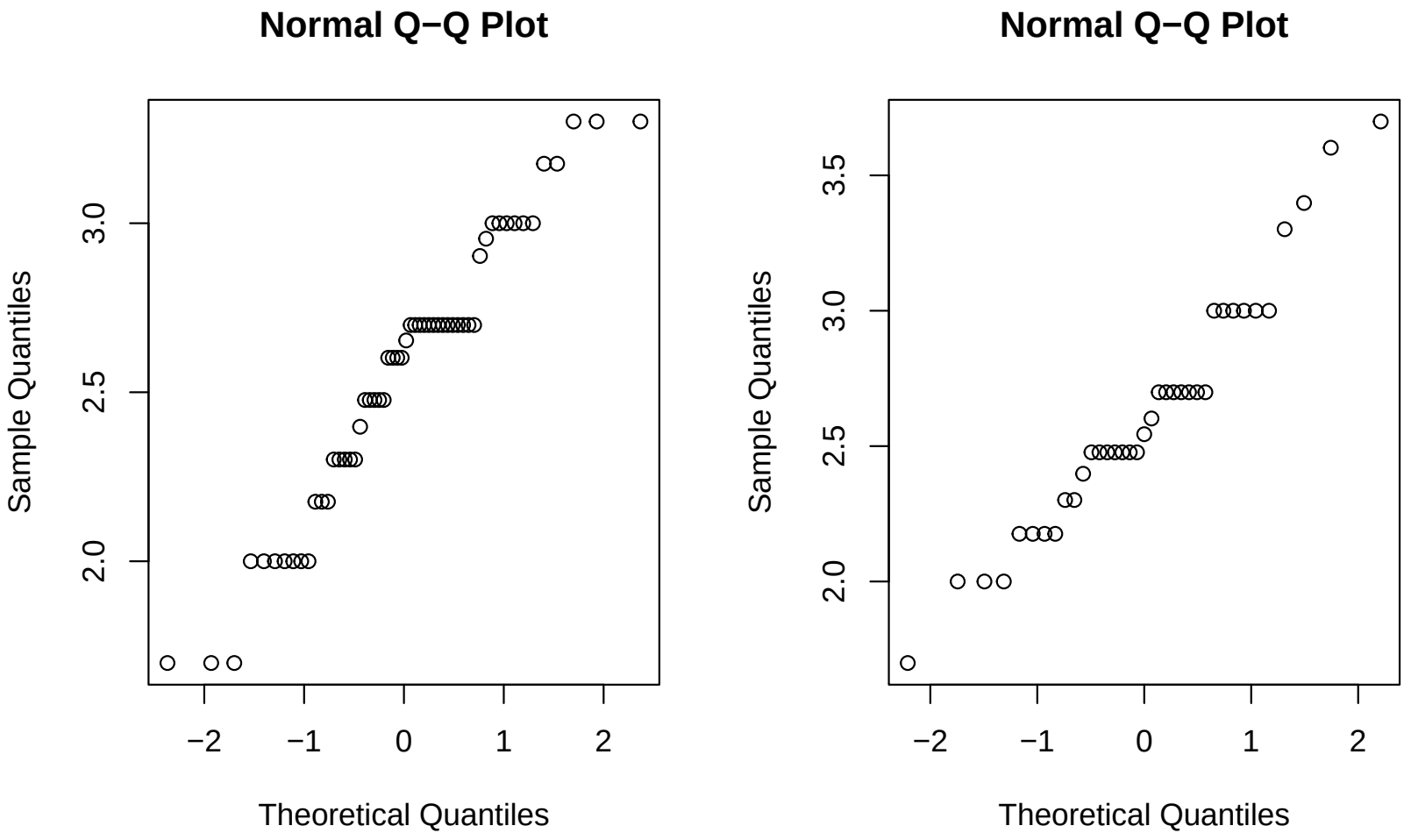


Diagramme des quantiles-quantiles

```
par(mfrow = c(1,2))
tapply(log10(d$Q12coutsoins), d$Q8bourse, qqnorm)
```



Test possible au vu des distributions mais dont on n'attend pas grand chose vu la proximité des 2 distributions

```
t.test(log10(d$Q12coutsoins) ~ d$Q8bourse,
       var.equal = FALSE)
```

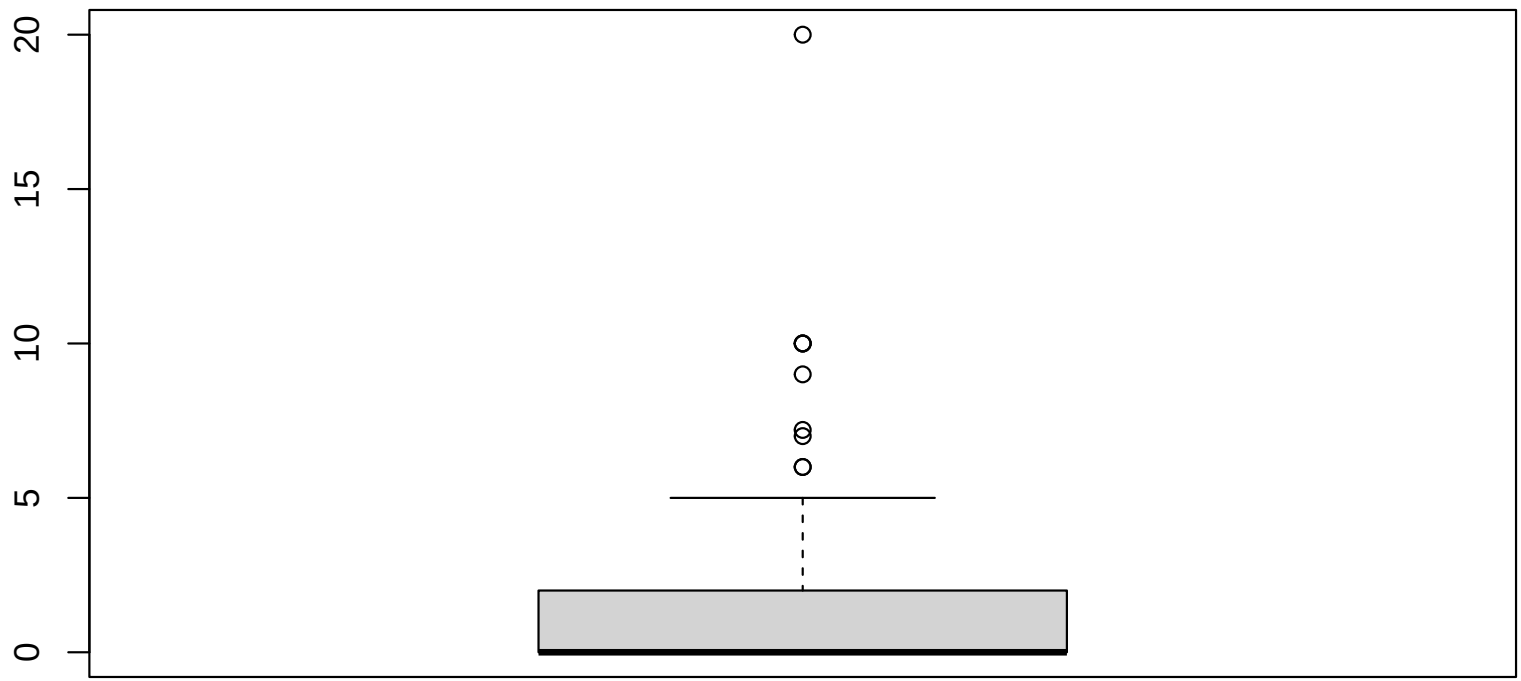
```
##
##  Welch Two Sample t-test
##
## data:  log10(d$Q12coutsoins) by d$Q8bourse
## t = -0.7, df = 72, p-value = 0.5
## alternative hypothesis: true difference in means betwee
## 95 percent confidence interval:
##  -0.248  0.120
## sample estimates:
## mean in group non mean in group oui
##              2.55              2.62
```

Cas d'une variable quantitative contenant une information qualitative

Distribution du nombre d'heures de travail rémunéré par semaine

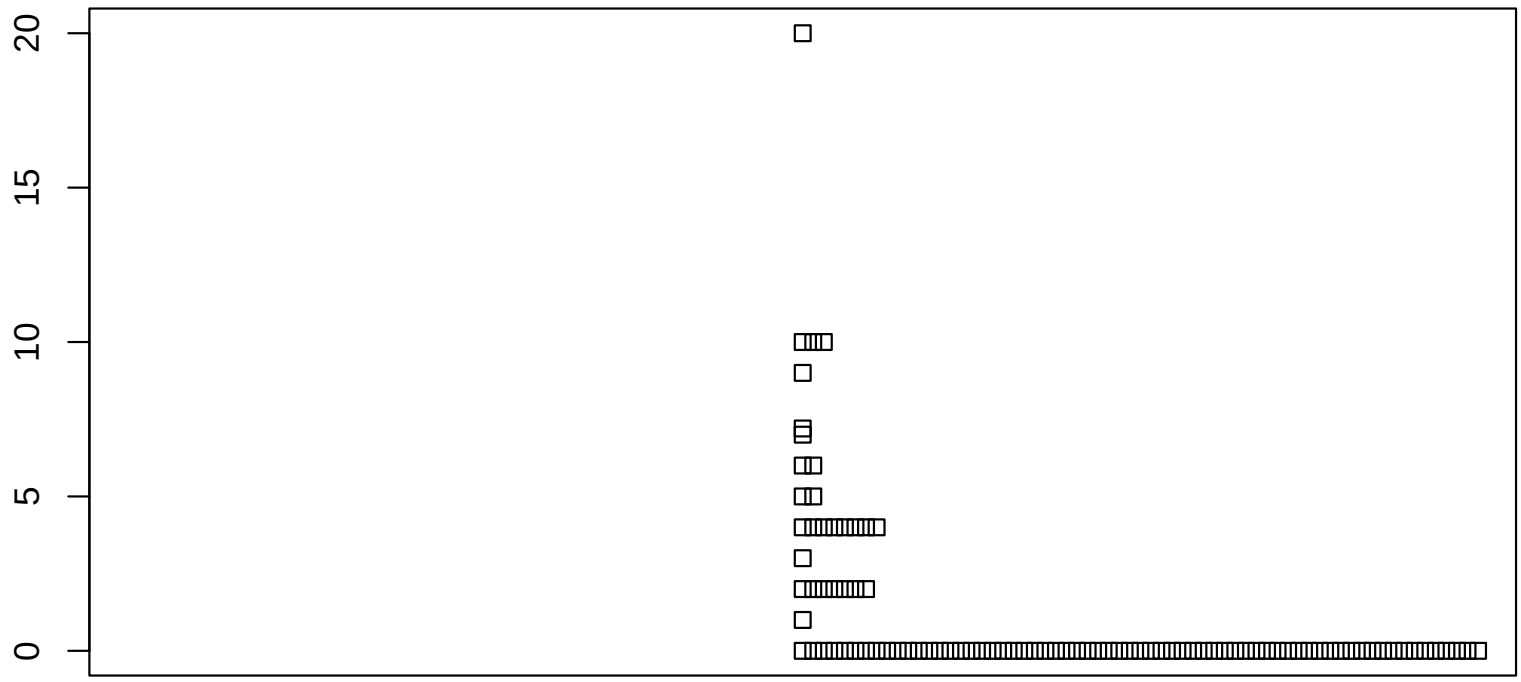
Q9 : Combien d'heures rémunérées travaillez-vous en moyenne par semaine (job étudiant, 0 si vous n'en avez pas) ?

```
boxplot(d$Q9job)
```



Regardons de plus près l'ensemble des points

```
stripchart(d$Q9job, vertical = TRUE, method = "stack")
```



Recodage possible en variable qualitative

Recodage simple en une variable à deux modalités (job : oui ou non).

```
d$job <- factor(d$Q9job > 0)
(table(d$job))
```

```
##
## FALSE  TRUE
##    65    28
```

Corrélation avec la question Q8 (bénéficiaire de bourse)

```
(t <- table(d$Q8bourse, d$job))
```

```
##
##        FALSE TRUE
##   non     43   13
##   oui     22   15
```

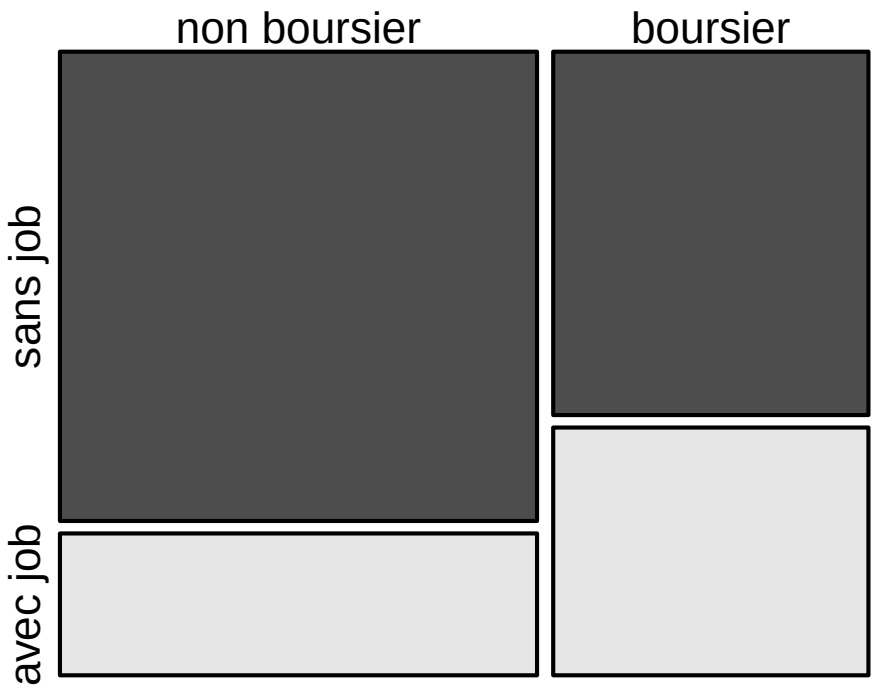
Renommons les modalités pour mieux s'y retrouver

```
levels(d$job) <- c("sans job", "avec job")
levels(d$Q8bourse) <- c("non boursier", "boursier")
(t <- table(d$Q8bourse, d$job))
```

```
##
##                  sans job avec job
##   non boursier         43       13
##   boursier             22       15
```

Visualisation de la table

```
plot(t, main = "", col = TRUE)
```



Calcul des fréquences d'étudiants avec et sans job parmi les boursiers et les non boursiers

```
prop.table(t, margin = 1)
```

```
##
##                  sans job avec job
##   non boursier      0.768    0.232
##   boursier          0.595    0.405
```

Tests possibles (1)

```
chisq.test(t)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity corre
##
## data:  t
## X-squared = 2, df = 1, p-value = 0.1
```

Tests possibles (2)

Equivalent avec en prime l'IC à 95% sur la diférence entre les fréquences (ici de sans job)

```
prop.test(t) # équivalent possible sur une table 2*2
```

```
##
##  2-sample test for equality of proportions with continu
##
## data:  t
## X-squared = 2, df = 1, p-value = 0.1
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0422  0.3887
## sample estimates:
## prop 1 prop 2
##  0.768  0.595
```

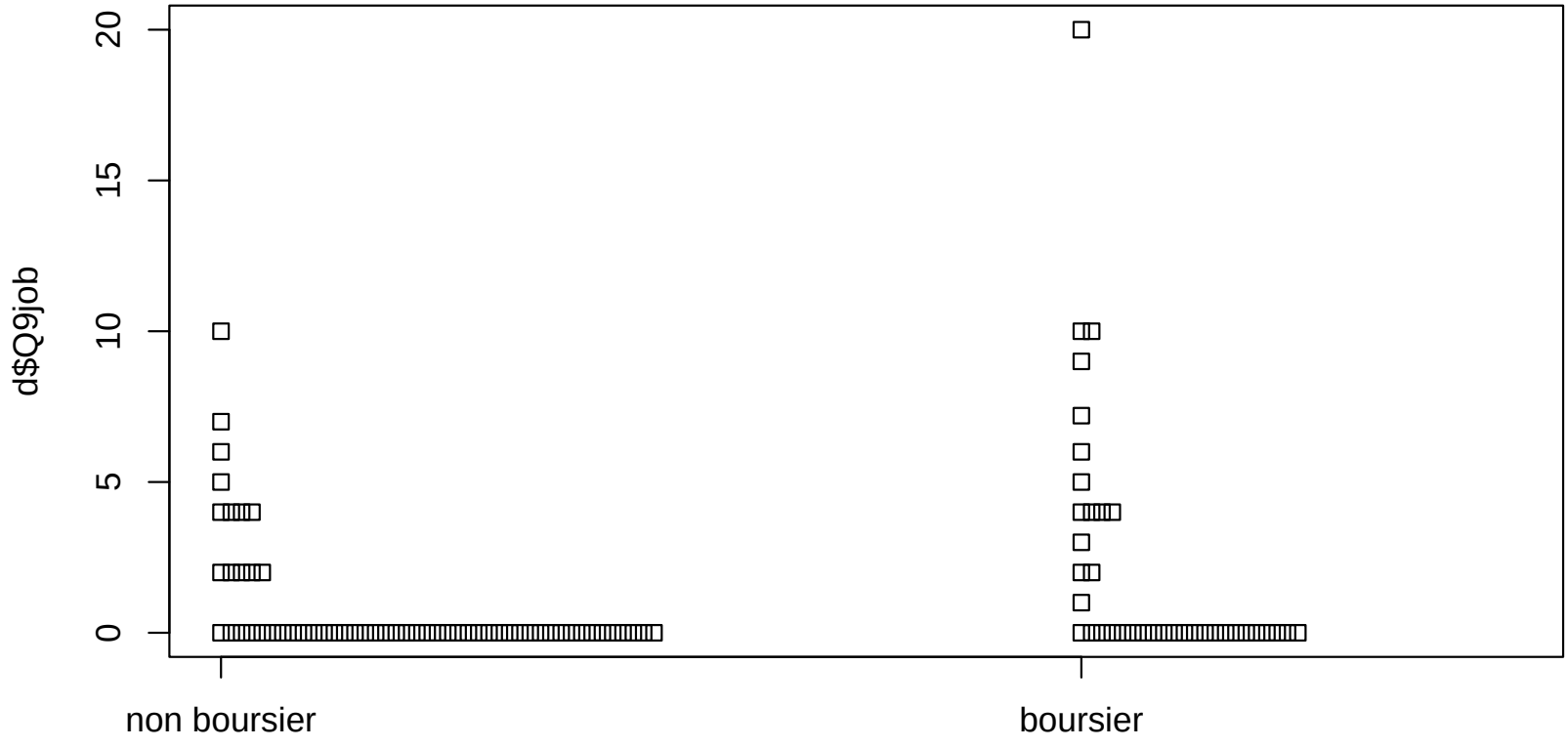
Tests possibles (3)

```
fisher.test(t) # test exact possible sur une table  2*2
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.1
## alternative hypothesis: true odds ratio is not equal to
## 95 percent confidence interval:
##  0.829 6.136
## sample estimates:
## odds ratio
##       2.23
```

Corrélation avec la question Q8 en gardant Q9 quantitative

```
stripchart(d$Q9job ~ d$Q8bourse, vertical = TRUE,
           method = "stack")
```



Test de la somme des rangs possible en gardant Q9 quantitative

```
wilcox.test(d$Q9job ~ d$Q8bourse, paired = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d$Q9job by d$Q8bourse
## W = 834, p-value = 0.05
## alternative hypothesis: true location shift is not equal
```

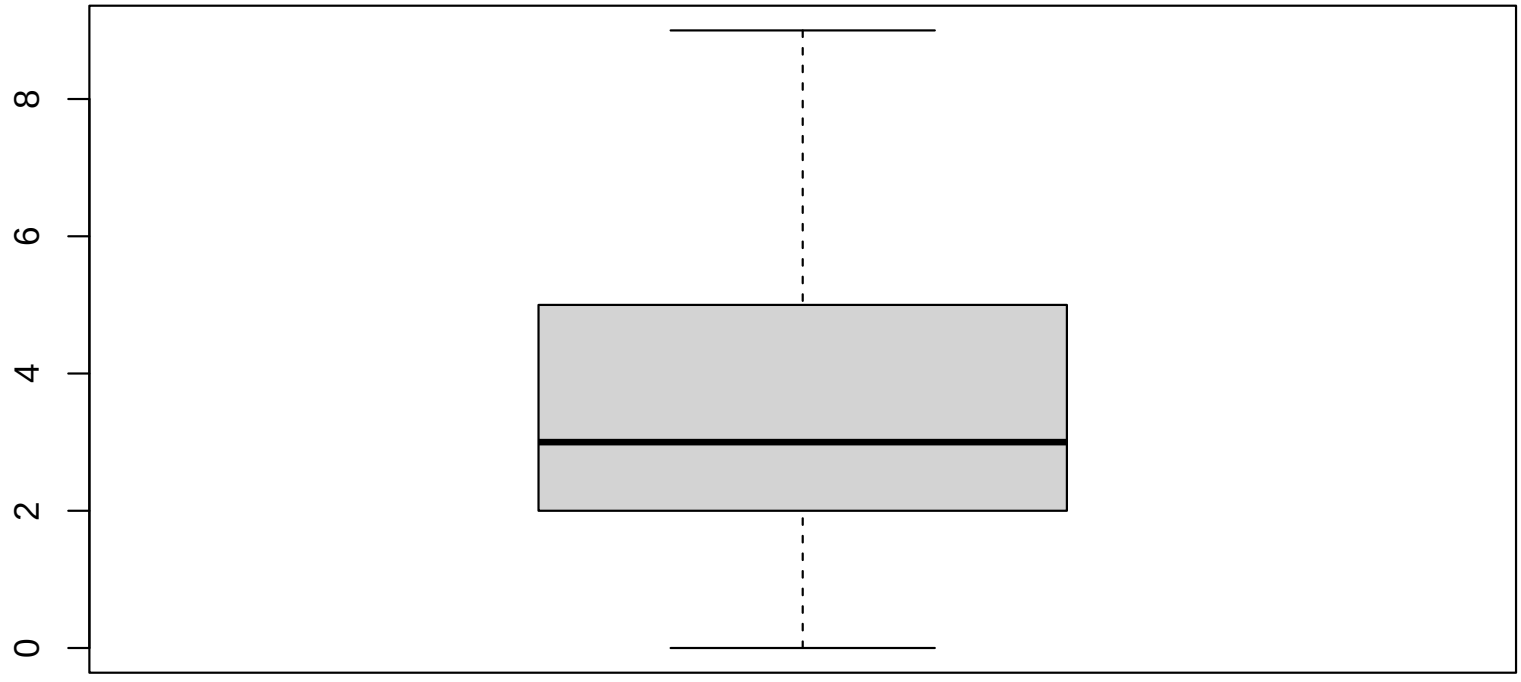
Il ne serait par contre pas raisonnable d'utiliser un test paramétrique et/ou de résumer les données par des moyennes.

Cas des variables quantitatives discrètes (similaires aux scores manipulés classiquement en clinique)

Distribution du nombre de clubs (gamme de 0 à 9)

Q21 : A combien de clubs participez-vous sur le campus ?

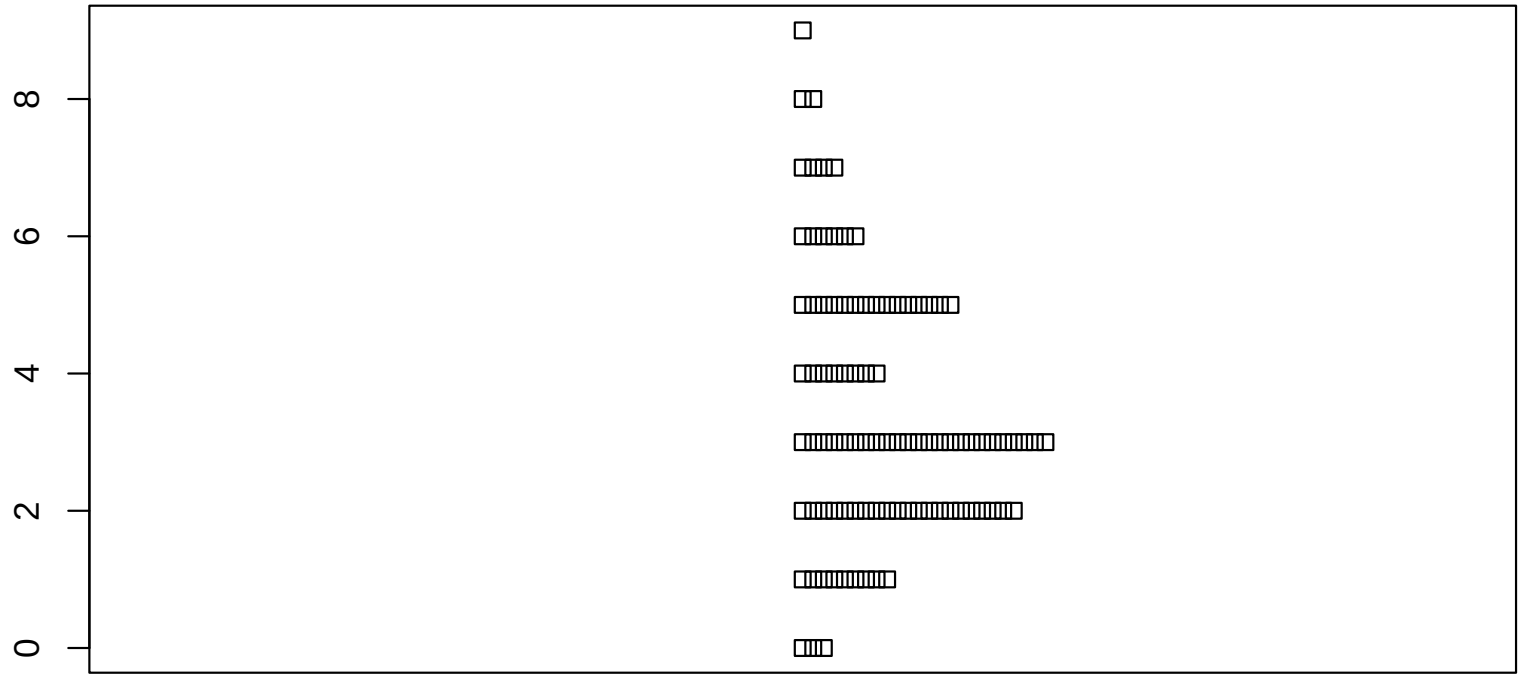
```
boxplot(d$Q21clubs)
```



Regardons de plus près en visualisant tous les points

Pas si loin d'une distribution normale !

```
stripchart(d$Q21clubs, vertical = TRUE, method = "stack")
```



Corrélation avec Q10 (animal de compagnie)

Q10: Vivez-vous actuellement avec un animal de compagnie ?

```
stripchart(d$Q21clubs ~ d$Q10animal, vertical = TRUE,
           method = "stack")
```

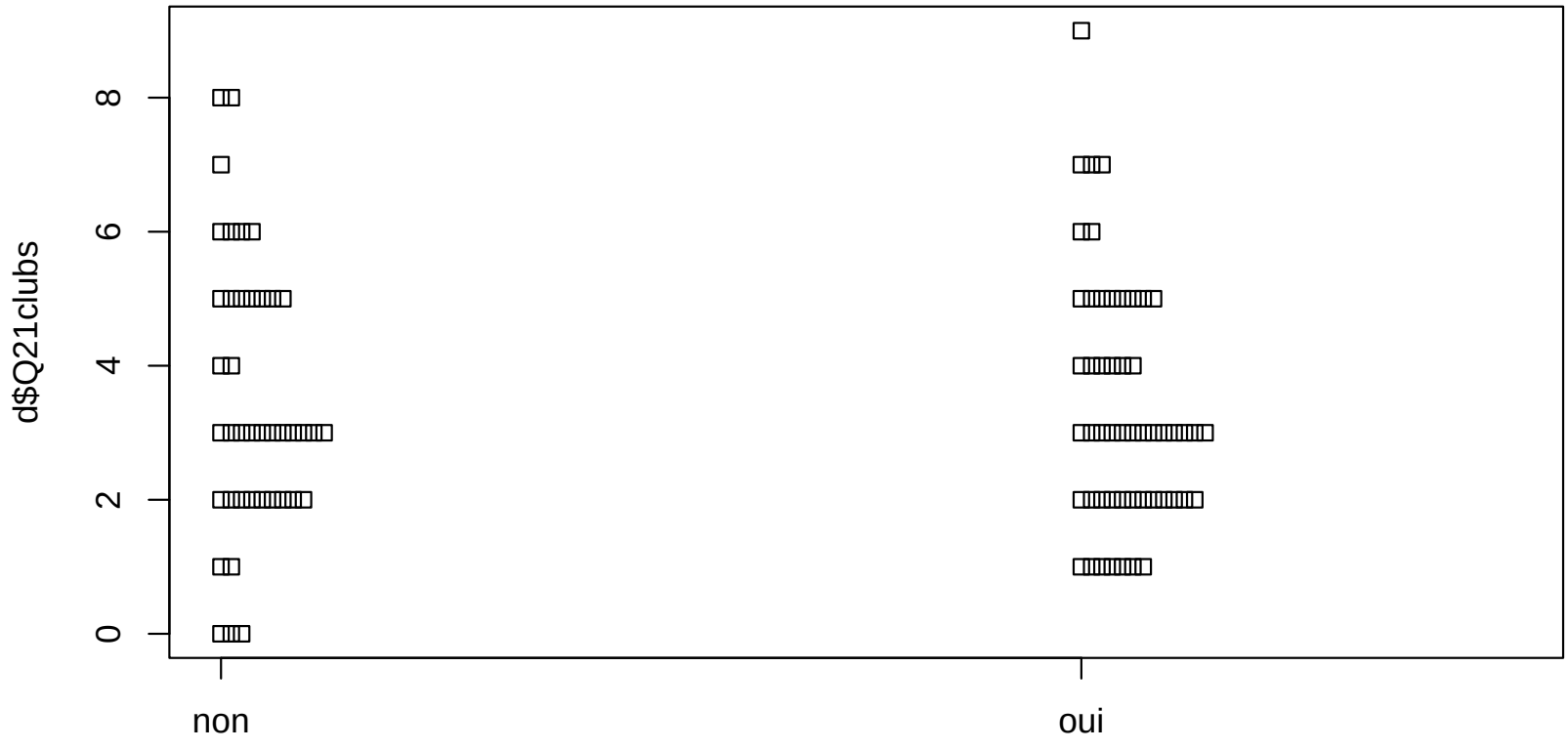
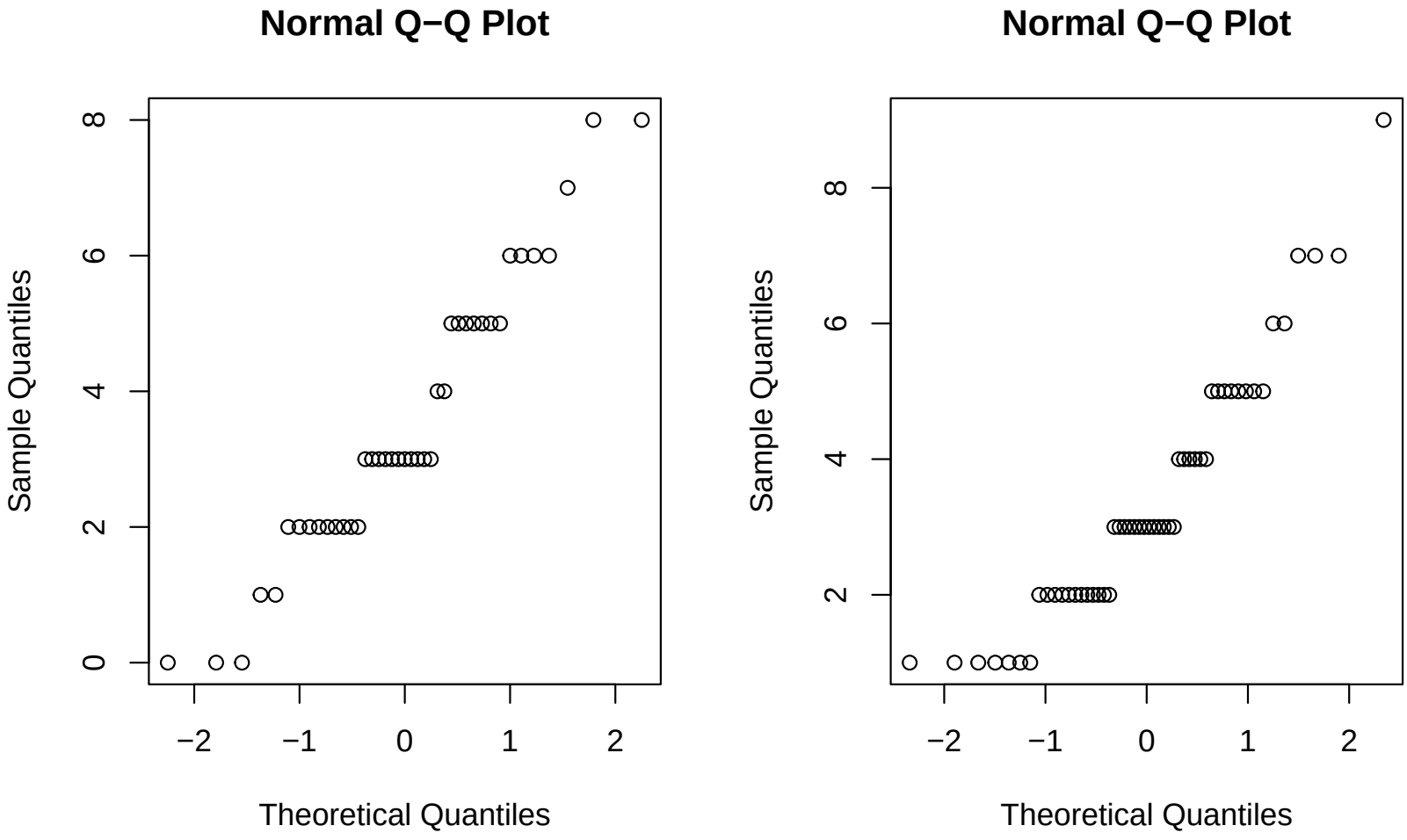


Diagramme des quantiles-quantiles

```
par(mfrow = c(1,2))
tapply(d$Q21clubs, d$Q10animal, qqnorm)
```



Une démarche paramétrique serait acceptable mais on n'en attend pas plus que l'estimation des moyennes et de leur différence vu la proximité des 2 distributions

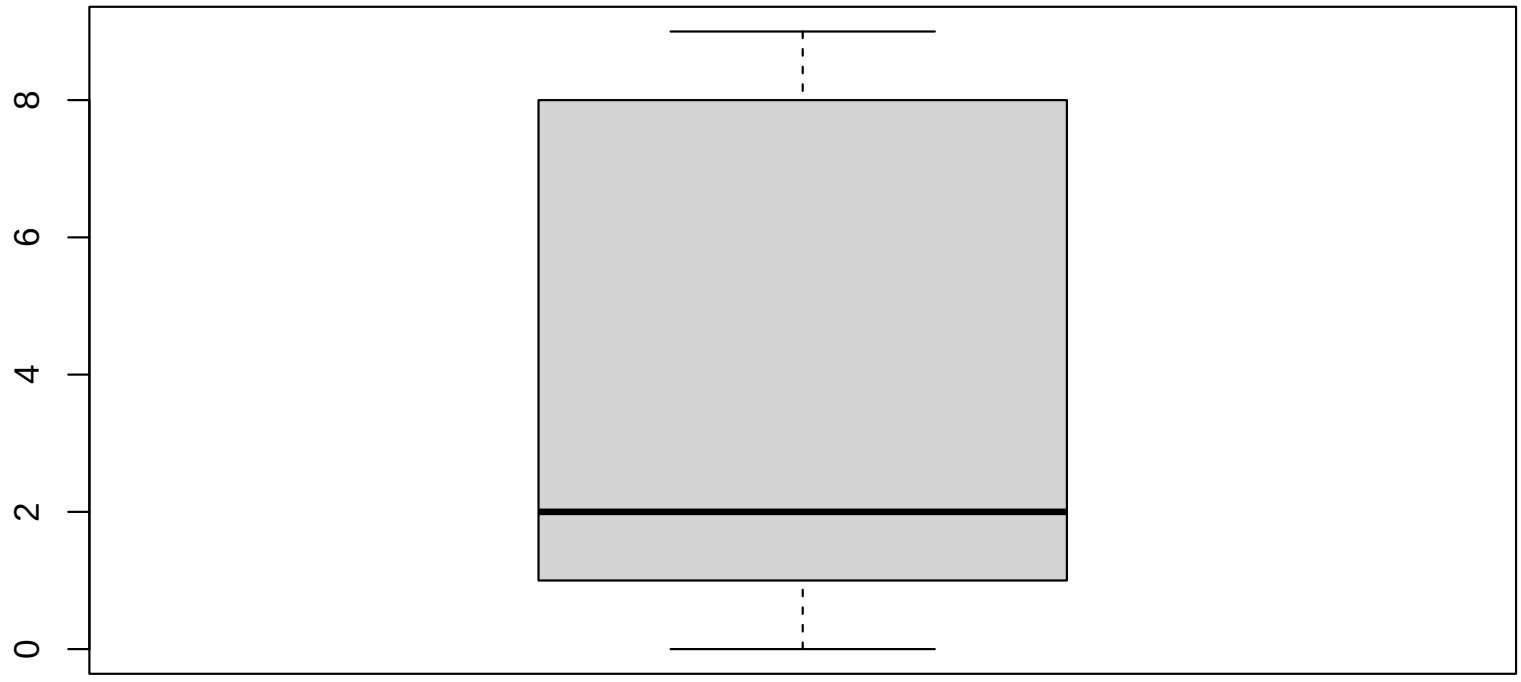
```
t.test(d$Q21clubs ~ d$Q10animal,
       var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  d$Q21clubs by d$Q10animal
## t = 0.3, df = 91, p-value = 0.8
## alternative hypothesis: true difference in means betwee
## 95 percent confidence interval:
##  -0.697  0.903
## sample estimates:
## mean in group non mean in group oui
##              3.49              3.38
```

Distribution du nb. cours suivis en amphi (de 0 à 9)

Q22 : Quel est le nombre d'unités d'enseignement pour lesquelles vous avez suivi au moins 75% des cours en amphi au second semestre (entre 0 et 9) ?

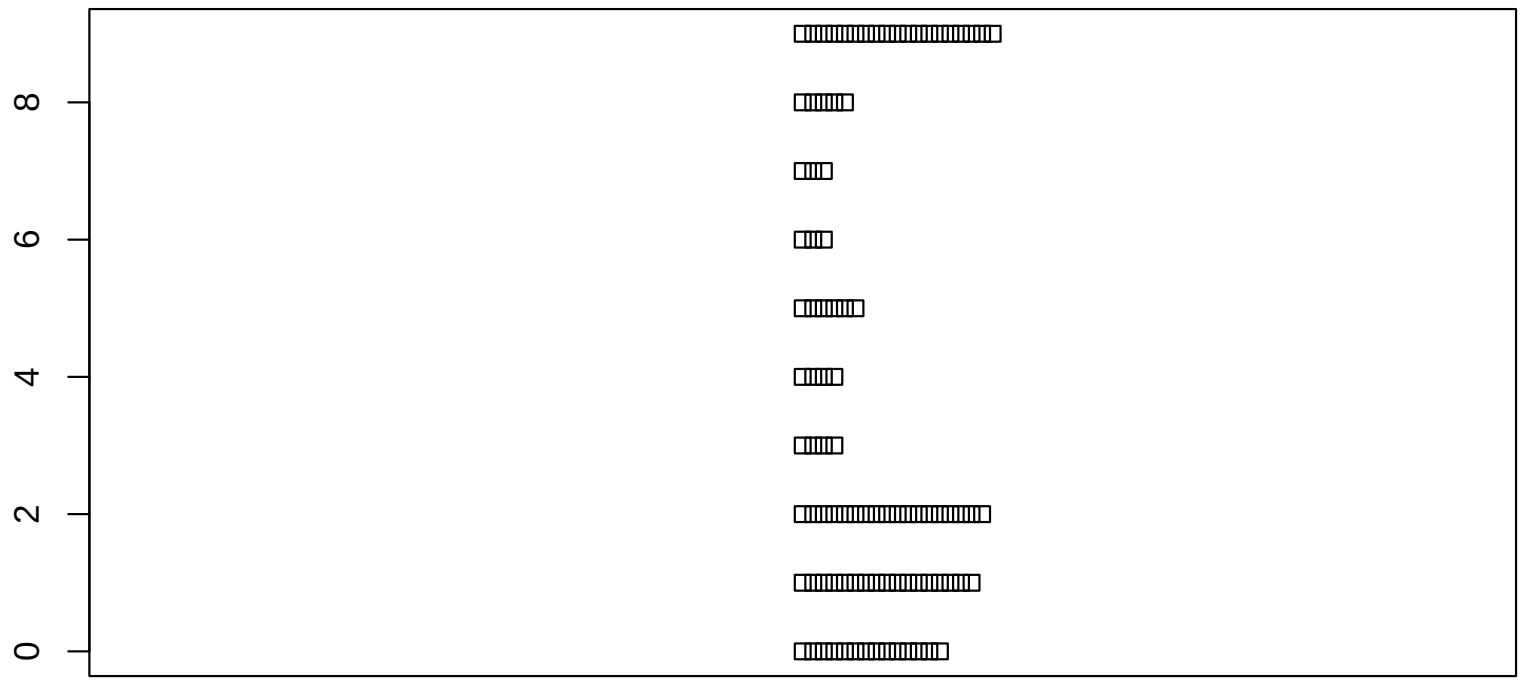
```
boxplot(d$Q22amphi)
```



Regardons de plus près en visualisant tous les points

On a beau avoir le même nombre de valeurs possibles (0 à 9), on ne traitera pas cette variable bimodale comme la précédente.

```
stripchart(d$Q22amphi, vertical = TRUE, method = "stack")
```



Définition d'une variable qualitative

Variable à deux classes

```
d$plus4amphi <- factor(d$Q22amphi > 4)
table(d$plus4amphi)
```

```
##
## FALSE  TRUE
##    57    36
```

Variable à plus de deux classes

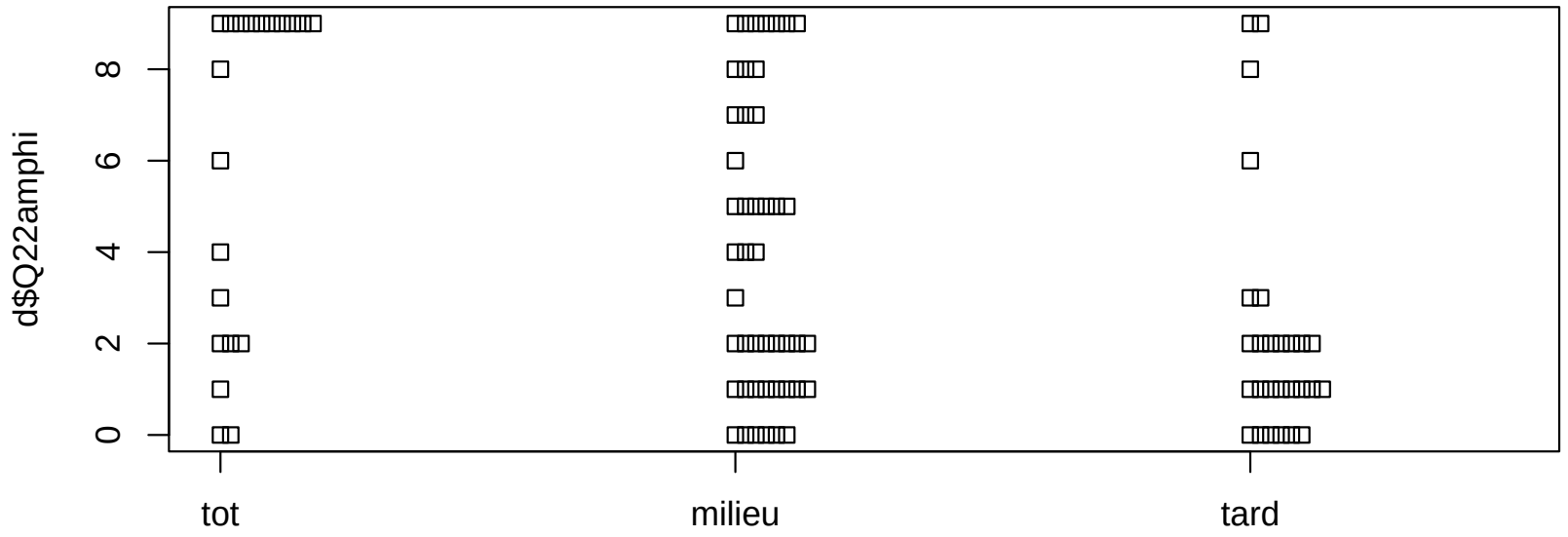
```
d$amphi <- cut(d$Q22amphi, breaks = c(0, 2, 6, 9))
table(d$amphi)
```

```
##
## (0,2] (2,6] (6,9]
##    35    17    27
```

Corrélation avec Q26

Q26: Quand vous mettez-vous à apprendre vos cours ? (le plus tôt possible, quelque temps après leur sortie, en catastrophe 3 jours avant les partiels)

```
d$Q26polys <- factor(d$Q26polys, # réordonnancement
                     levels = c("tot", "milieu", "tard"))
stripchart(d$Q22amphi ~ d$Q26polys, vertical = TRUE, metho
```

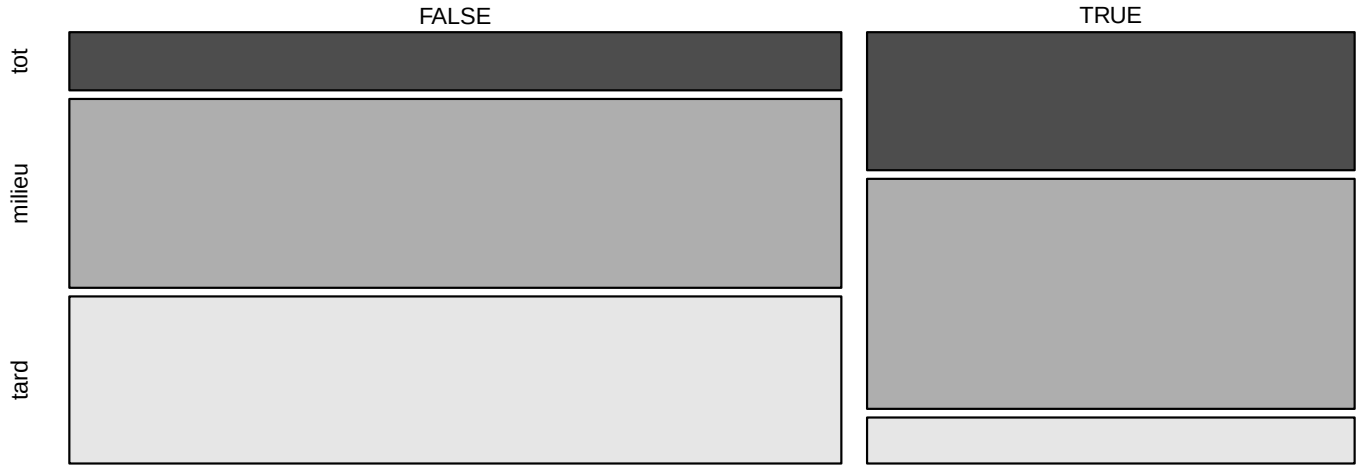


Corrélation avec Q26 (Q22 en variable à deux modalités)

```
(t <- table(d$plus4amphi, d$Q26polys))
```

```
##
##           tot milieu tard
##    FALSE    8      26    23
##    TRUE    12      20     4
```

```
plot(t, main = "", col = TRUE)
```



Test réalisable (montrant une corrélation significative)

```
(khi2 <- chisq.test(t))
```

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 11, df = 2, p-value = 0.005
```

```
khi2$expected # vérification des conditions d'utilisation
```

```
##
##           tot milieu tard
##   FALSE 12.26   28.2 16.5
##   TRUE   7.74   17.8 10.5
```

Test des rangs (possible aussi) et mettant aussi en évidence une corrélation significative dans ce cas

```
kruskal.test(d$Q22amphi ~ d$Q26polys)
```

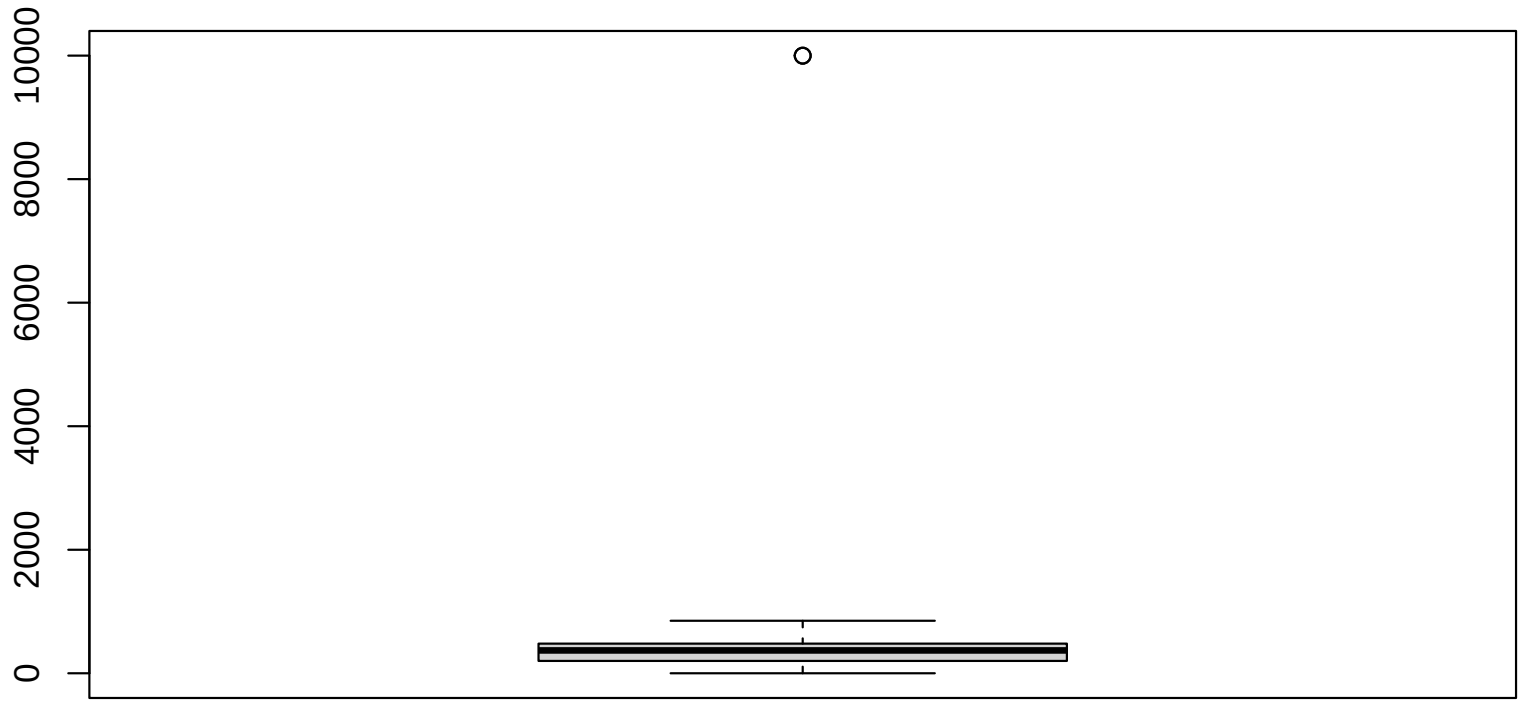
```
##
##  Kruskal-Wallis rank sum test
##
## data:  d$Q22amphi by d$Q26polys
## Kruskal-Wallis chi-squared = 13, df = 2, p-value = 0.002
```

Cas d'une variable quantitative avec des valeurs extrêmes

Distribution de la distance entre domiciles personnel et parental

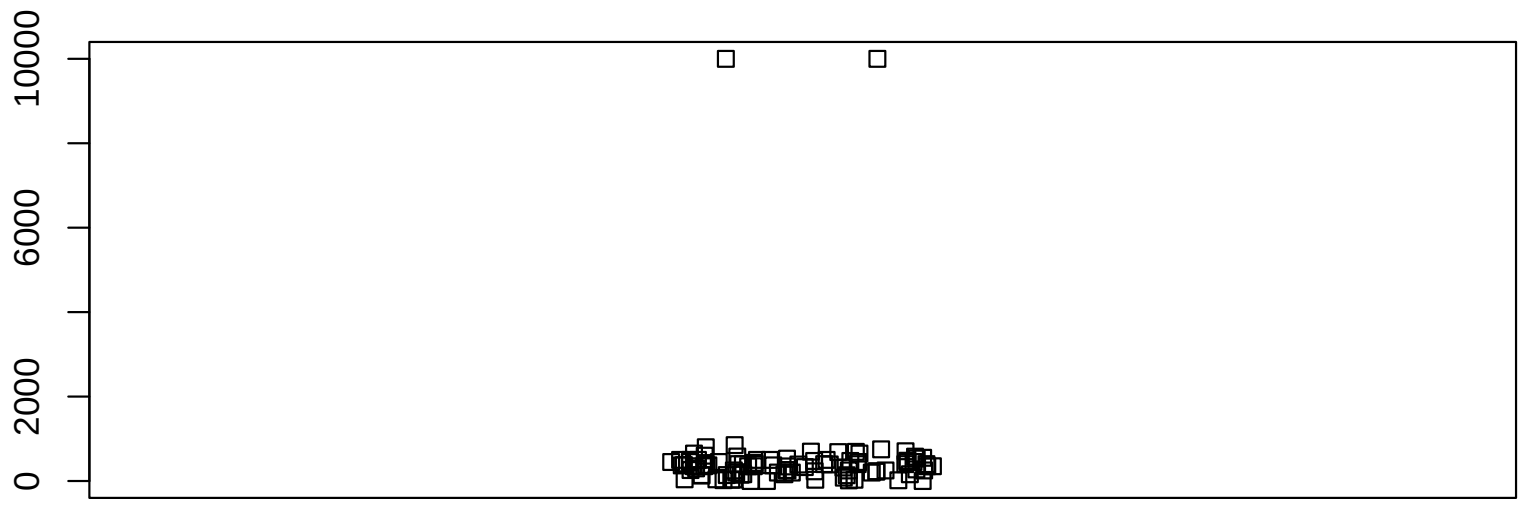
Q24 : Quelle est la distance entre votre domicile personnel et le domicile parental (en km) ?

```
boxplot(d$Q24km)
```



Regardons de plus près l'ensemble des points

```
stripchart(d$Q24km, vertical = TRUE, method = "jitter")
```



Les valeurs à 9999 sont en fait plus grandes mais on n'avait pas donné la possibilité de mettre plus de chiffres (erreur !).

Que faire dans un tel cas ?

Transformation logarithmique ?

Difficile dans un tel cas car on a des vrais 0.

```
which(d$Q24km == 0) # numéros des lignes avec 0 pour Q24km
```

```
## [1] 14 25 42
```

Définition de classes ?

Possible mais le choix des classes n'est pas trivial.

Utilisation de méthodes paramétriques sans transformation ?

Surtout pas. Trop peu robustes par rapport aux valeurs extrêmes !

Utilisation de méthodes non paramétriques ?

Démarche raisonnable mais on perd le côté descriptif.

Petite conclusion

Que faire avec des variables semi-quantitatives ?

Il n'y a pas de règle.

Tout dépend de la forme de la distribution.

- Il est important de regarder les points sous forme de "stripchart" surtout pour des variables quantitatives discrètes.
- Si un test doit être utilisé il doit être adapté à la forme des distributions observées.
- La moyenne n'est pas forcément un bon descripteur des données : elle résume parfois très mal les tendances observées et est très peu robuste par rapport aux valeurs extrêmes.

Le jeu de données pris en exemple est un jeu de données "jouet" et il ne conviendrait pas d'analyser les données d'une enquête structurée de cette façon en corrélant deux à deux les variables (cf. introduction aux méthodes multivariées).