TP3

Thomas Mauran

2023-04-04

TP-3 Everest

Liens du gif (latex n'accepte pas les gifs): https://media.tenor.com/g9bG3yk53YEAAAAM/god-of-war-god-of-war1.gif

Nombres d'enfants

Individu: Un couple

Population: Les couples francais

Variable: le nombre d'enfant de moins de 25 ans par couples. variable quantitative discrète

Les modalités: au nombre d'enfants de moins de 25 ans par couple, allant de zéro à un nombre élevé.

```
library(ggforce)
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
theme_set(theme_light())
library(here)
```

```
## here() starts at /home/thomas/Desktop/DO/R/do3-dataviz/RenduTP
```

```
library (readr)
library(tidyr)

couples <- read_delim("~/Desktop/DO/R/do3-dataviz/RenduTP/data/rp2017_td_fam2.csv",
    delim = "\t", escape_double = FALSE,
    col_types = cols(...8 = col_skip()),
    trim_ws = TRUE, skip = 6)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...8`
```

```r
couples <- na.omit(couples)
colnames(couples) <- c("Situation", "0", "1", "2", "3", "4", "total")
couples$Situation <- c("mariées", "pacsees", "concubinage", "autre", "total")
head(couples)
```

```
## # A tibble: 5 x 7
##   Situation       `0`     `1`     `2`     `3`    `4`    total
##   <chr>         <dbl>   <dbl>   <dbl>   <dbl>  <dbl>    <dbl>
## 1 mariées     6448133 1644613 1975639  798166 263408 11129960
## 2 pacsees      407144  337083  335833   62577  11225  1153862
## 3 concubinage 1304386  673141  564489  167676  61358  2771049
## 4 autre        177322   53585   41589   17237   8778   298511
## 5 total       8336985 2708422 2917549 1045657 344769 15353382
```

Décrivez les données en quelques mots.

Les données donnent le nombre d'enfants de moins de 25 ans par couple en France en 2017. La majorité des couples ont aucun ou un ou deux enfants de moins de 25 ans. Les couples mariés ont en moyenne plus d'enfants que les autres types de couples. Les couples pacsés ont en moyenne moins d'enfants que les couples mariés. Les couples ayant un autre statut conjugal ont en moyenne le moins d'enfants.

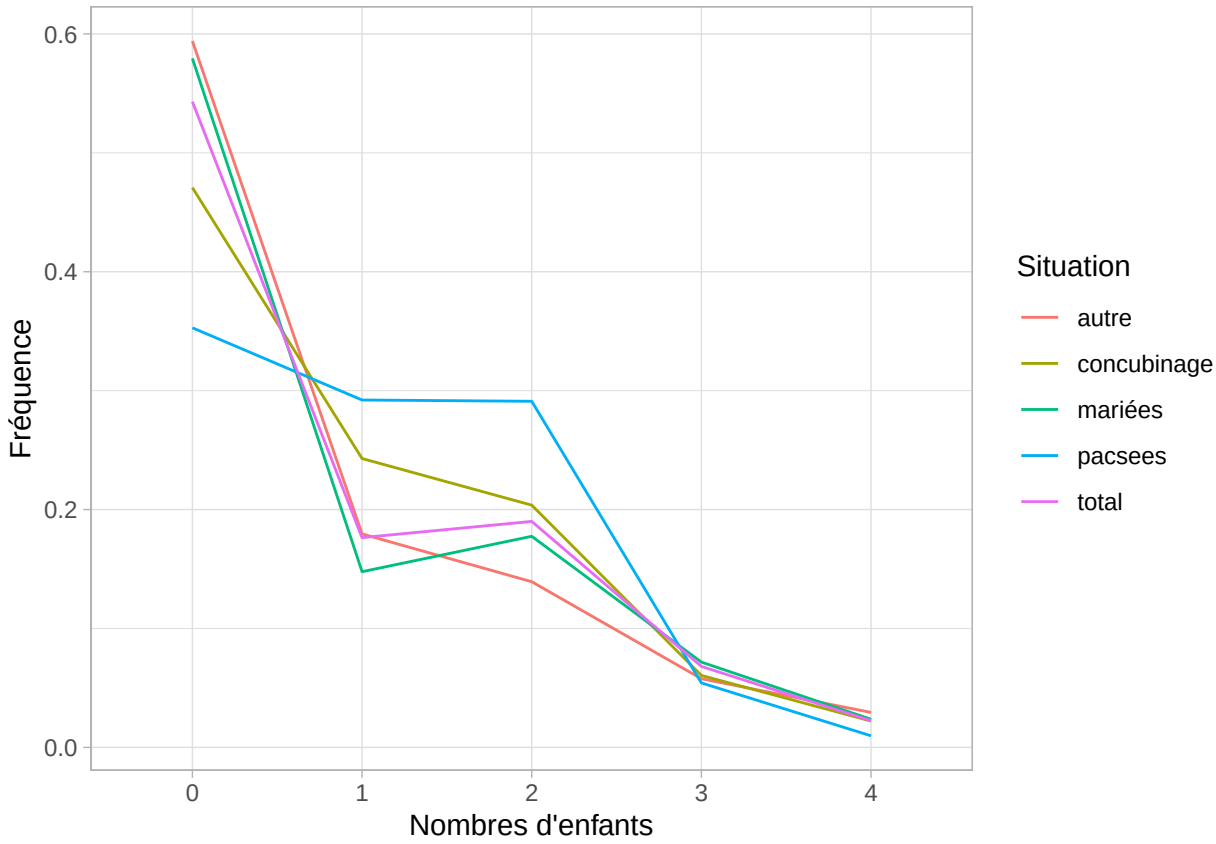
Formatage

```r
formattedTable <- pivot_longer(couples, cols = c ("0", "1", "2", "3", "4"), names_to="enfants", values_
formattedTable
```

```
## # A tibble: 25 x 4
##    Situation    total enfants  compte
##    <chr>        <dbl> <chr>     <dbl>
##  1 mariées   11129960 0       6448133
##  2 mariées   11129960 1       1644613
##  3 mariées   11129960 2       1975639
##  4 mariées   11129960 3        798166
##  5 mariées   11129960 4        263408
##  6 pacsees    1153862 0        407144
##  7 pacsees    1153862 1        337083
##  8 pacsees    1153862 2        335833
##  9 pacsees    1153862 3         62577
## 10 pacsees    1153862 4         11225
## # i 15 more rows
```

Polygone de fréquence

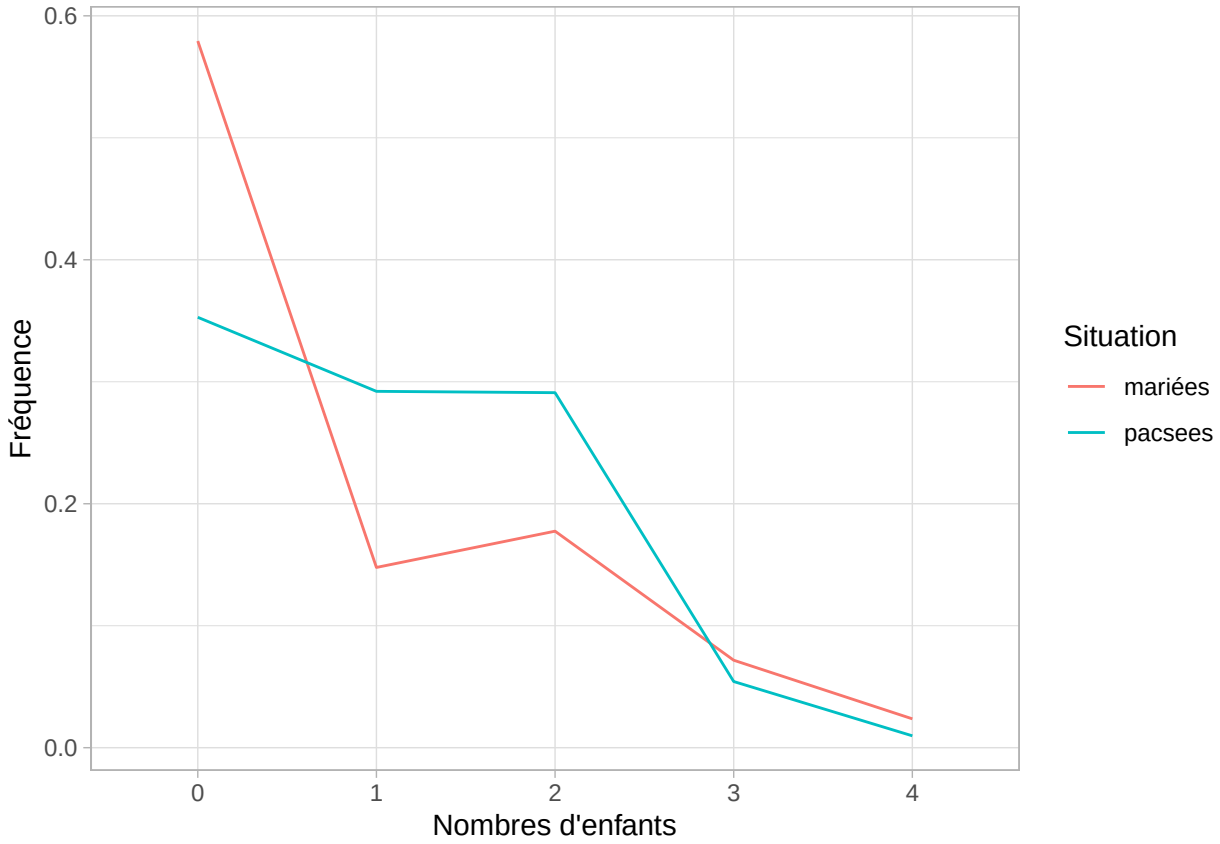
```r
ggplot(formattedTable, aes(x = enfants,
                           y = compte / total,
                           color = Situation,
                           group = Situation))+
  geom_line() +
  xlab("Nombres d'enfants") +
  ylab("Fréquence")
```



Ici une representation en ligne nous permet facilement de nous rendre compte de la répartition du nombre d'enfant en fonction de la situation du couple. Ici on se rend compte que les mariés ont plus souvent 0 enfants que les gens pacsees. Les couples pacsees ont tendance à avoir plus souvent 1 ou 2 enfants que les couples mariées. Les couples mariés ont en revanche plus souvent 3 ou 4 enfants que les couples pacsees

Graphe simplifié

```
subset <- subset(formattedTable, Situation == "mariées" | Situation == "pacsees" )
ggplot(subset, aes(x = enfants,
                         y = compte / total,
                         color = Situation,
                         group = Situation))+
  geom_line() +
  xlab("Nombres d'enfants") +
  ylab("Fréquence")
```



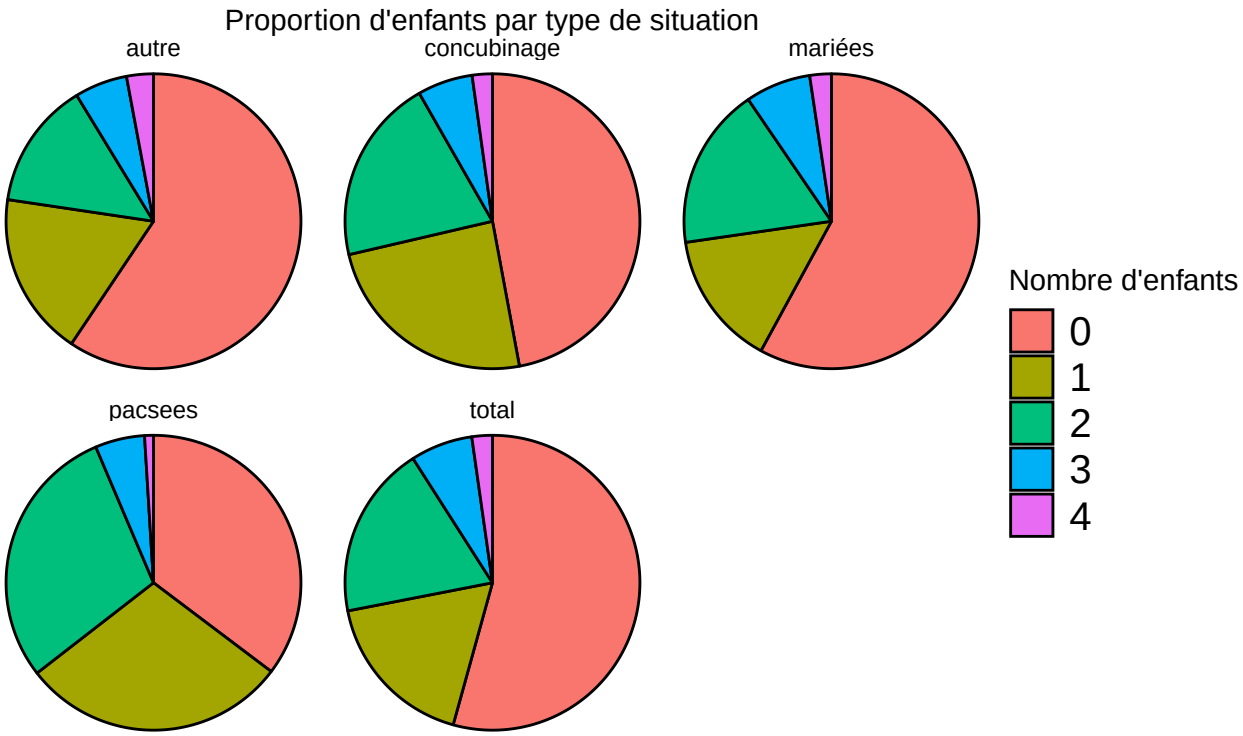
ce graphe nous permet de mieux nous rendre compte des différences de fréquences du nombre d'enfants par couple pacsees ou mariés

Autre question

Comment se répartit le nombre d'enfants selon les différentes situations étudiées?

```
subsetAutreQuestion <- subset(formattedTable)
pie <- ggplot(subsetAutreQuestion,
              aes(x0 = 0, y0 = 0, r0 = 0, r = 1,
                  amount = compte / total,
                  fill = as.factor(enfants))) +
  coord_fixed() +
  ggtitle("") +
  geom_arc_bar(stat = "pie") +
  ggtitle("Proportion d'enfants par type de situation") +
  labs( fill = "Nombre d'enfants") +
  theme_void() +
  theme(plot.title = element_text(size = 11, hjust = 0.5),
        legend.position = "right",
        legend.text = element_text(size = 15),
        axis.title = element_blank()) +
    facet_wrap(~Situation, ncol = 3)

pie
```



Everest

```
expeditions <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
```

```
## ‘curl‘ package not installed, falling back to using ‘url()‘
## Rows: 76519 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): expedition_id, member_id, peak_id, peak_name, season, sex, citizen...
## dbl  (5): year, age, highpoint_metres, death_height_metres, injury_height_me...
## lgl  (6): hired, success, solo, oxygen_used, died, injured
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
expeditions
```

```
## # A tibble: 76,519 x 21
##    expedition_id member_id    peak_id peak_name   year season sex     age
##    <chr>         <chr>        <chr>   <chr>      <dbl> <chr>  <chr> <dbl>
##  1 AMAD78301     AMAD78301-01 AMAD    Ama Dablam  1978 Autumn M        40
##  2 AMAD78301     AMAD78301-02 AMAD    Ama Dablam  1978 Autumn M        41
##  3 AMAD78301     AMAD78301-03 AMAD    Ama Dablam  1978 Autumn M        27
##  4 AMAD78301     AMAD78301-04 AMAD    Ama Dablam  1978 Autumn M        40
##  5 AMAD78301     AMAD78301-05 AMAD    Ama Dablam  1978 Autumn M        34
##  6 AMAD78301     AMAD78301-06 AMAD    Ama Dablam  1978 Autumn M        25
##  7 AMAD78301     AMAD78301-07 AMAD    Ama Dablam  1978 Autumn M        41
```

```
##  8 AMAD78301     AMAD78301-08 AMAD    Ama Dablam  1978 Autumn M        29
##  9 AMAD79101     AMAD79101-03 AMAD    Ama Dablam  1979 Spring M        35
## 10 AMAD79101     AMAD79101-04 AMAD    Ama Dablam  1979 Spring M        37
## # i 76,509 more rows
## # i 13 more variables: citizenship <chr>, expedition_role <chr>, hired <lgl>,
## #   highpoint_metres <dbl>, success <lgl>, solo <lgl>, oxygen_used <lgl>,
## #   died <lgl>, death_cause <chr>, death_height_metres <dbl>, injured <lgl>,
## #   injury_type <chr>, injury_height_metres <dbl>
```

Description des données

Les données fournies comprennent des informations sur des expéditions d'alpinisme. Chaque ligne représente un membre de l'expédition et contient des détails tels que le nom de l'expédition, le membre de l'expédition, le nom et l'ID du pic, l'année et la saison de l'expédition, le sexe, l'âge, la nationalité, le rôle dans l'expédition, l'embauche, l'altitude du sommet atteint, le succès de l'expédition, si elle a été effectuée en solo, l'utilisation d'oxygène, la mort, la cause du décès, la hauteur du décès, les blessures, le type de blessure et la hauteur de la blessure.

Description de l'experience statistique

Question : "Comment se répartit l'âge des membres d'une expédition réussie vers le Mont Everest ?"

Individu: Les Alpinistes du Mont Everest

Population: l'ensemble des membres des expéditions de l'Everest

Échantillon: Les membres ayant réussit l'expedition

Variable: L'age des membres

Modalité: Un nombre

Sélectionnez dans le tableau uniquement les lignes répondant à ces critères, et dont l'âge n'est pas manquant.

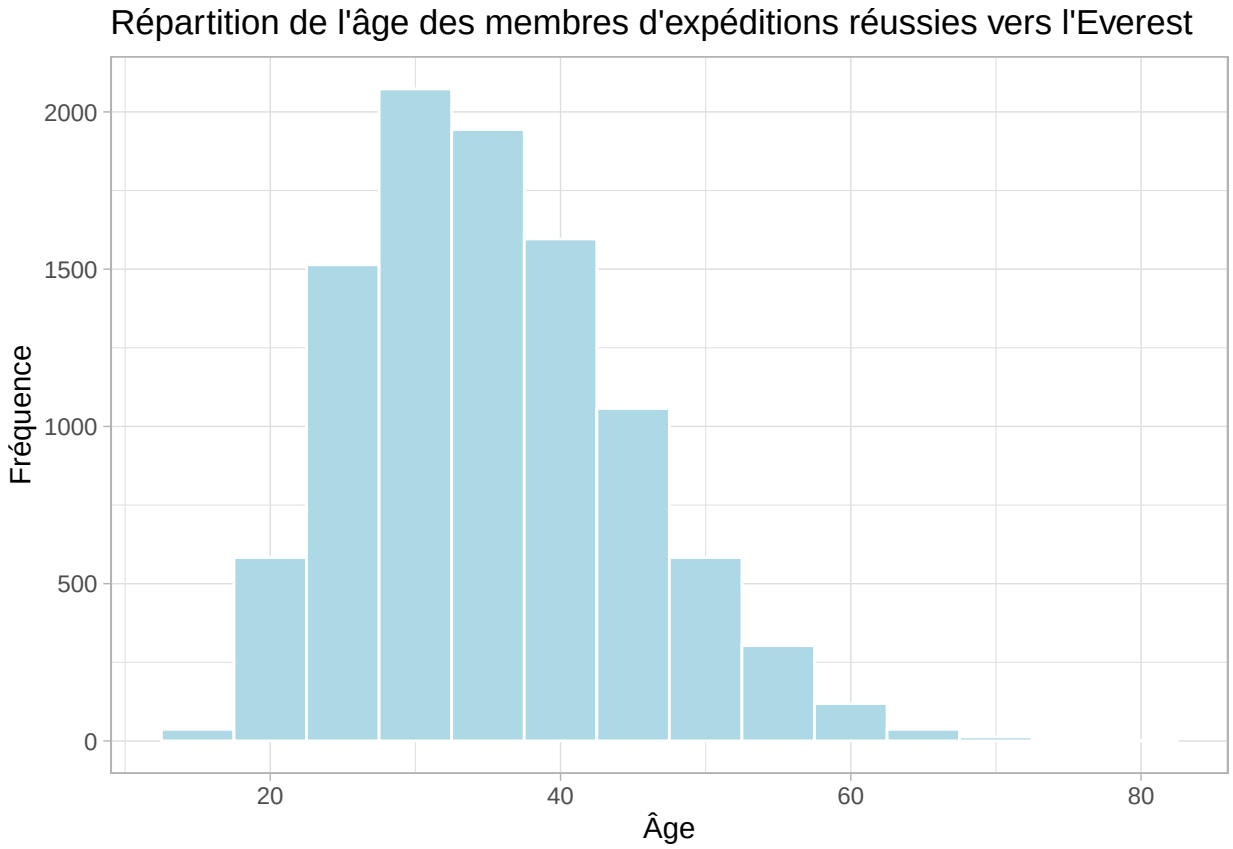
```
everest <- subset(expeditions, success==TRUE & peak_name=="Everest" & !is.na(age))
everest
```

```
## # A tibble: 9,863 x 21
##    expedition_id member_id    peak_id peak_name  year season sex     age
##    <chr>         <chr>        <chr>   <chr>     <dbl> <chr>  <chr> <dbl>
##  1 EVER63101     EVER63101-04 EVER    Everest    1963 Spring M        31
##  2 EVER63101     EVER63101-10 EVER    Everest    1963 Spring M        32
##  3 EVER63101     EVER63101-11 EVER    Everest    1963 Spring M        26
##  4 EVER63101     EVER63101-19 EVER    Everest    1963 Spring M        36
##  5 EVER63101     EVER63101-20 EVER    Everest    1963 Spring M        34
##  6 EVER63101     EVER63101-21 EVER    Everest    1963 Spring M        26
##  7 EVER65101     EVER65101-03 EVER    Everest    1965 Spring M        28
##  8 EVER65101     EVER65101-11 EVER    Everest    1965 Spring M        27
##  9 EVER65101     EVER65101-04 EVER    Everest    1965 Spring M        42
## 10 EVER65101     EVER65101-05 EVER    Everest    1965 Spring M        23
## # i 9,853 more rows
## # i 13 more variables: citizenship <chr>, expedition_role <chr>, hired <lgl>,
## #   highpoint_metres <dbl>, success <lgl>, solo <lgl>, oxygen_used <lgl>,
## #   died <lgl>, death_cause <chr>, death_height_metres <dbl>, injured <lgl>,
## #   injury_type <chr>, injury_height_metres <dbl>
```

Représentez ces données sous la forme d'un histogramme. Justifiez le choix de la largeur des classes.

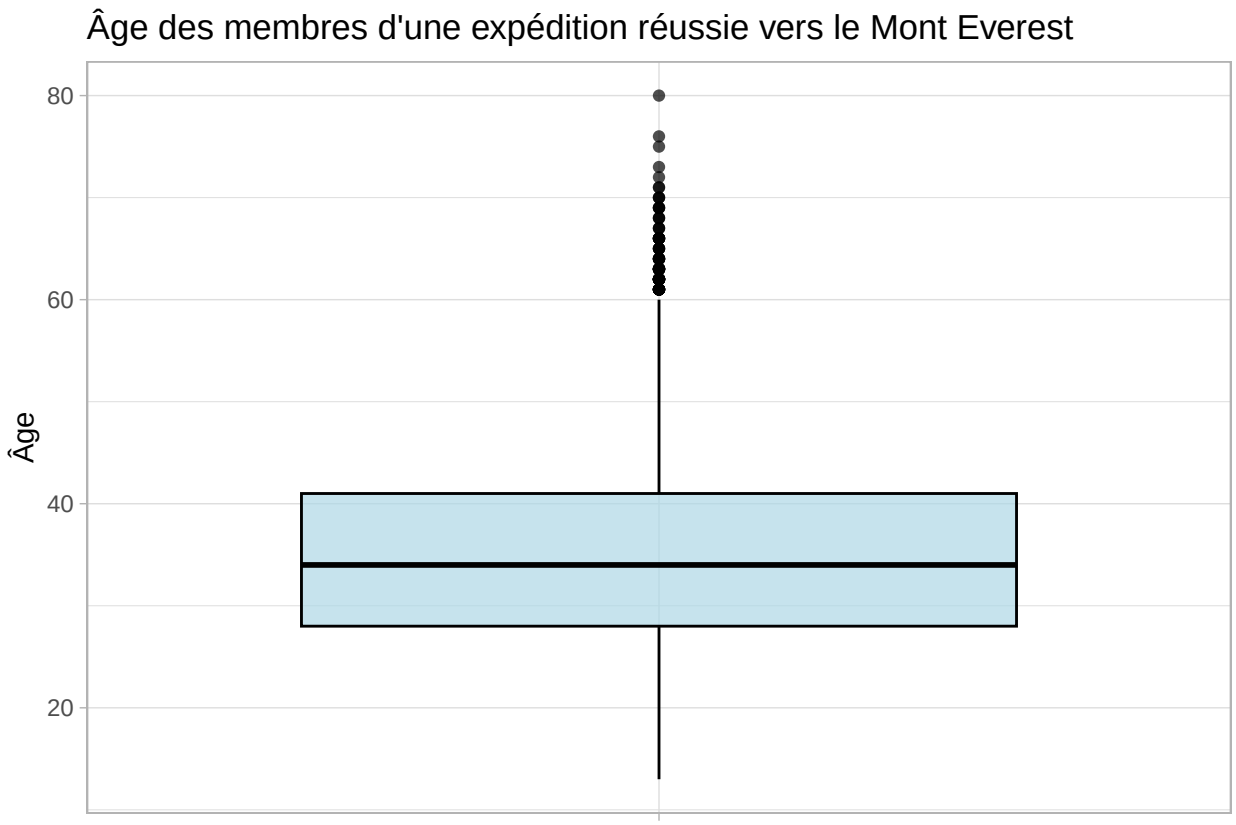
```r
library(ggplot2)

ggplot(data = everest,
       aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "white") +
  labs(title = "Répartition de l'âge des membres d'expéditions réussies vers l'Everest",
       x = "Âge",
       y = "Fréquence")
```



Représentez ces même données sous la forme d'une boite à moustache (boxplot).

```r
ggplot(everest, aes(x="", y=age)) +
  geom_boxplot(fill="lightblue", color="black", alpha=0.7) +
  labs(x=NULL, y="Âge") +
  ggtitle("Âge des membres d'une expédition réussie vers le Mont Everest")
```



Il est dit "**d´ une** boite à moustache' dans l'énoncé alors je l'ai représenté de la sorte même si le graphique me semble bizarre

Laquelle de ces représentations est la plus informative ? Justifiez.

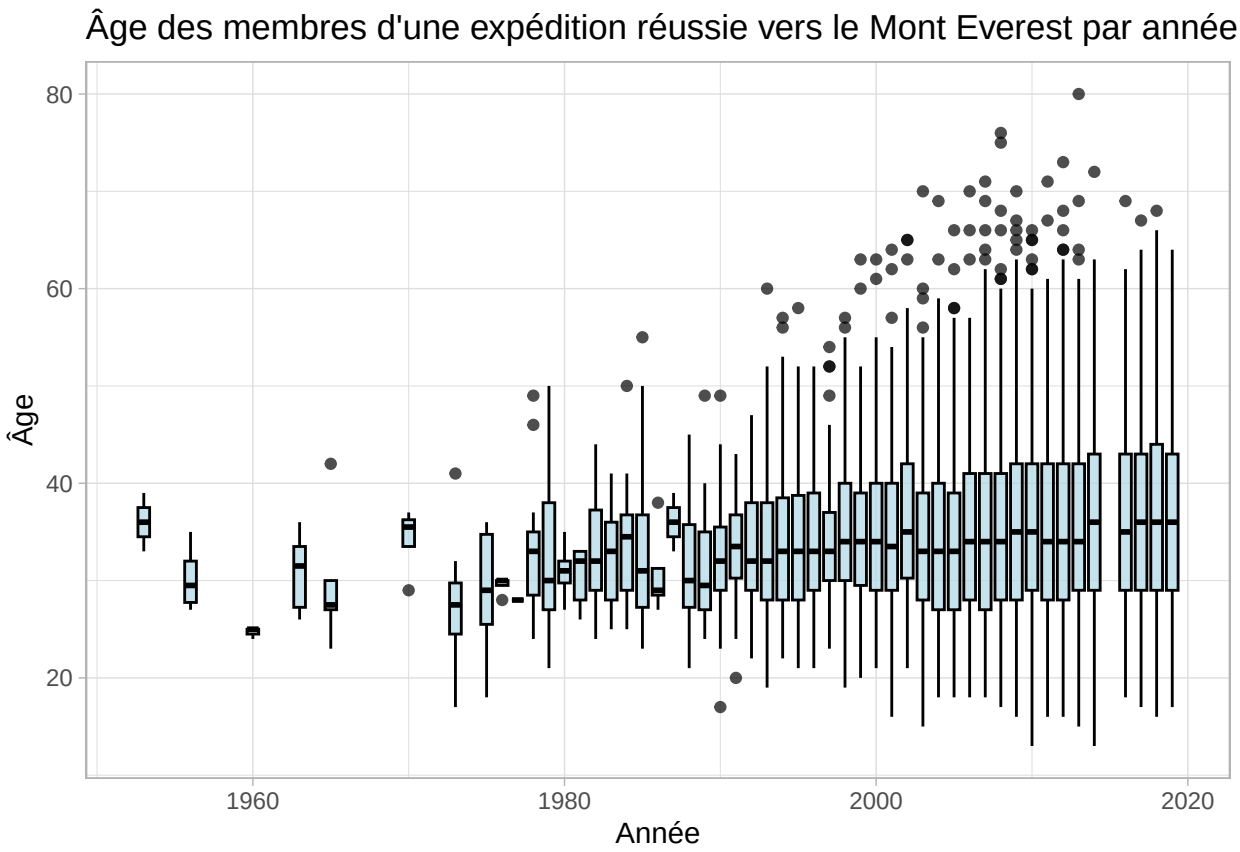
Les deux representations sont informatives. Je pense que l'histograme permet mieux de representer une répartition. En voyant les données étalées en plusieurs barres de différentes tailles on se rend bien compte d'ou se trouve la plupart des données. Néanmoins le diagramme en moustache permet de mieux remarquer les valeurs aberrantes.

Que pouvez-vous dire sur l'âge des membres d'une expédition réussie vers le Mont Everest ?

On remarque que la plupart du temps les membres qui réussissent une expédition vers le mont Everest ont entre 25 et 40 ans. On peut aussi noter la présence très rare de personne bien plus jeune ou bien plus vieille dans des expéditions réussite

Age en fonction des années d'ascension

```
ggplot(everest, aes(x=year, y=age, group=year)) +
  geom_boxplot(fill="lightblue", color="black", alpha=0.7, position=position_dodge(width=0.75)) +
  labs(x="Année", y="Âge") +
  ggtitle("Âge des membres d'une expédition réussie vers le Mont Everest par année")
```



Avec cette représentation on se rend bien compte de plusieurs choses. Au fil des années la répartition des âges des personnes qui monte leverest devient de plus en plus large. On peut aussi penser que de plus en plus de personnent montent l'Everest. On peut aussi remarquer que l'âge médian augmente d'année en année.

Age des membres d'une expédition réussie ou non

Question : "Y-a-t-il une différence d'âge entre les membres d'une expédition réussie, et ceux d'une expédition qui a échoué, avec ou sans oxygène ?"

Individus : Les membres des expéditions réussies et échouées du Mont Everest, avec et sans oxygène.

Population : L'ensemble des membres de toutes les expéditions du Mont Everest, passées et futures.

Échantillon : Les membres des expéditions réussies et échouées du Mont Everest, avec et sans oxygène, pour lesquels l'âge est connu.

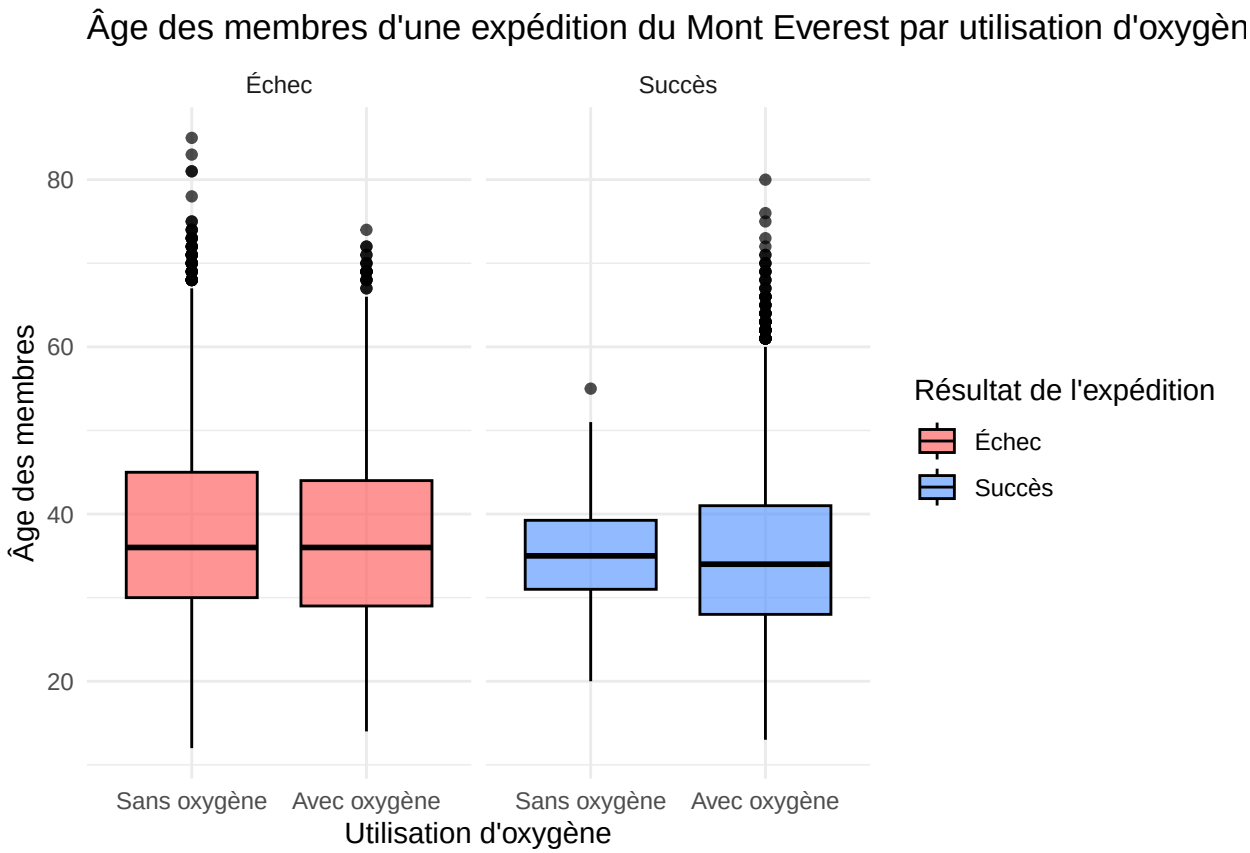
Variables : Le statut de l'expédition (réussie ou échouée), l'utilisation d'oxygène, l'âge des membres.

Modalités : Pour le statut de l'expédition, la modalité est soit "réussie" soit "échouée". Pour l'utilisation d'oxygène, la modalité est soit "avec oxygène" soit "sans oxygène". Pour l'âge des membres, la modalité est un nombre.

```
tab <- subset(expeditions, peak_name == "Everest" & !is.na(age))

ggplot(tab, aes(x = oxygen_used, y = age, fill = success)) +
  geom_boxplot(alpha = 0.7, color = "black") +
  facet_wrap(success ~ ., labeller = as_labeller(c(`TRUE` = "Succès", `FALSE` = "Échec"))) +
  labs(x = "Utilisation d'oxygène", y = "Âge des membres", fill = "Résultat de l'expédition") +
  ggtitle("Âge des membres d'une expédition du Mont Everest par utilisation d'oxygène et résultat") +
```

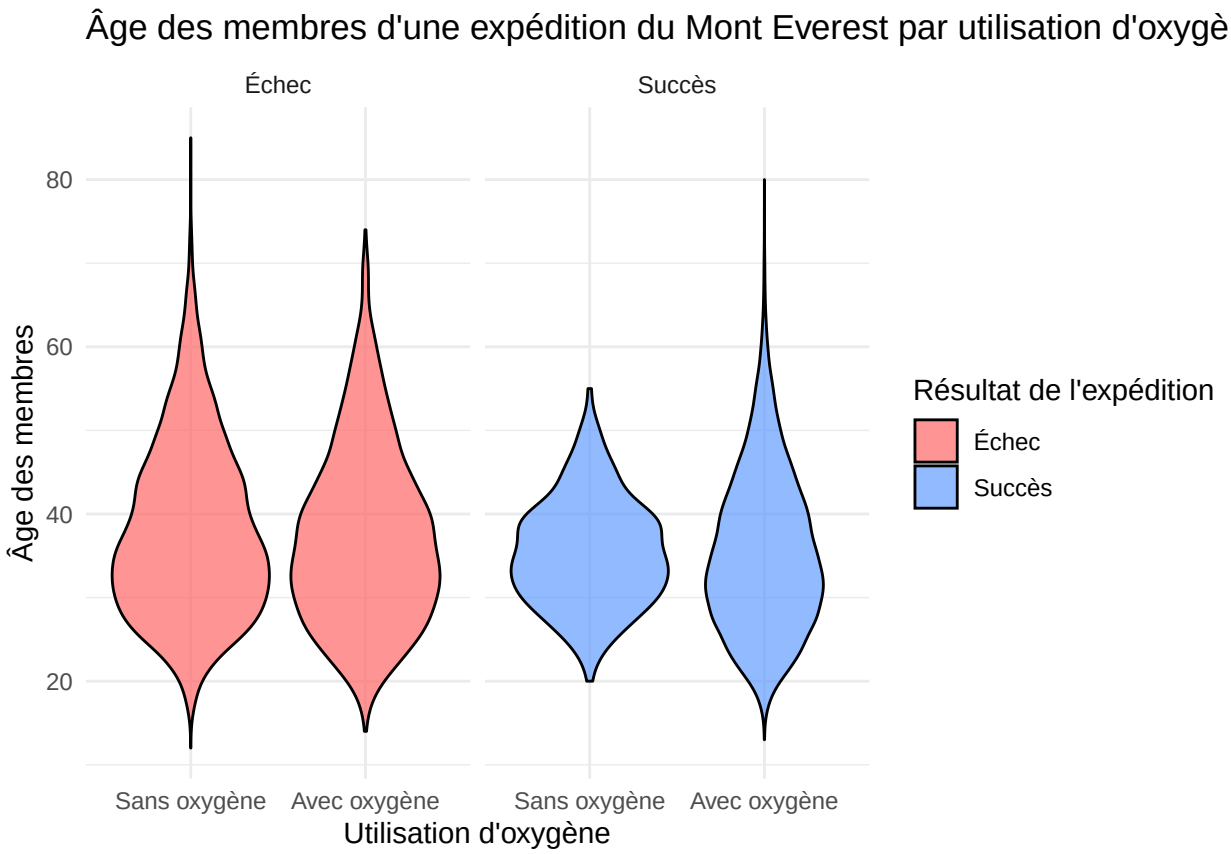
```
  scale_x_discrete(labels = c("Sans oxygène", "Avec oxygène")) +
  scale_fill_manual(values = c("#FF6666", "#619CFF"), labels = c("Échec", "Succès")) +
  theme_minimal()
```



On remarque que la distribution des âges est plus étalée dans le cas des ascensions réussies avec oxygène par rapport aux ascensions réussit sans oxygène.

```
tab <- subset(expeditions, peak_name == "Everest" & !is.na(age))

ggplot(tab, aes(x = oxygen_used, y = age, fill = success)) +
  geom_violin(alpha = 0.7, color = "black") +
  facet_wrap(success ~ ., labeller = as_labeller(c(`TRUE` = "Succès", `FALSE` = "Échec"))) +
  labs(x = "Utilisation d'oxygène", y = "Âge des membres", fill = "Résultat de l'expédition") +
  ggtitle("Âge des membres d'une expédition du Mont Everest par utilisation d'oxygène et résultat") +
  scale_x_discrete(labels = c("Sans oxygène", "Avec oxygène")) +
  scale_fill_manual(values = c("#FF6666", "#619CFF"), labels = c("Échec", "Succès")) +
  theme_minimal()
```



Autre question

Description de l'experience statistique

Question : "Y-a-t-il une différence d'âge entre les membres d'une expédition réussie, et ceux d'une expédition qui a échoué, avec ou sans équipe ?"

Individu: Les Alpiniste

Population: l'ensemble des alpiniste

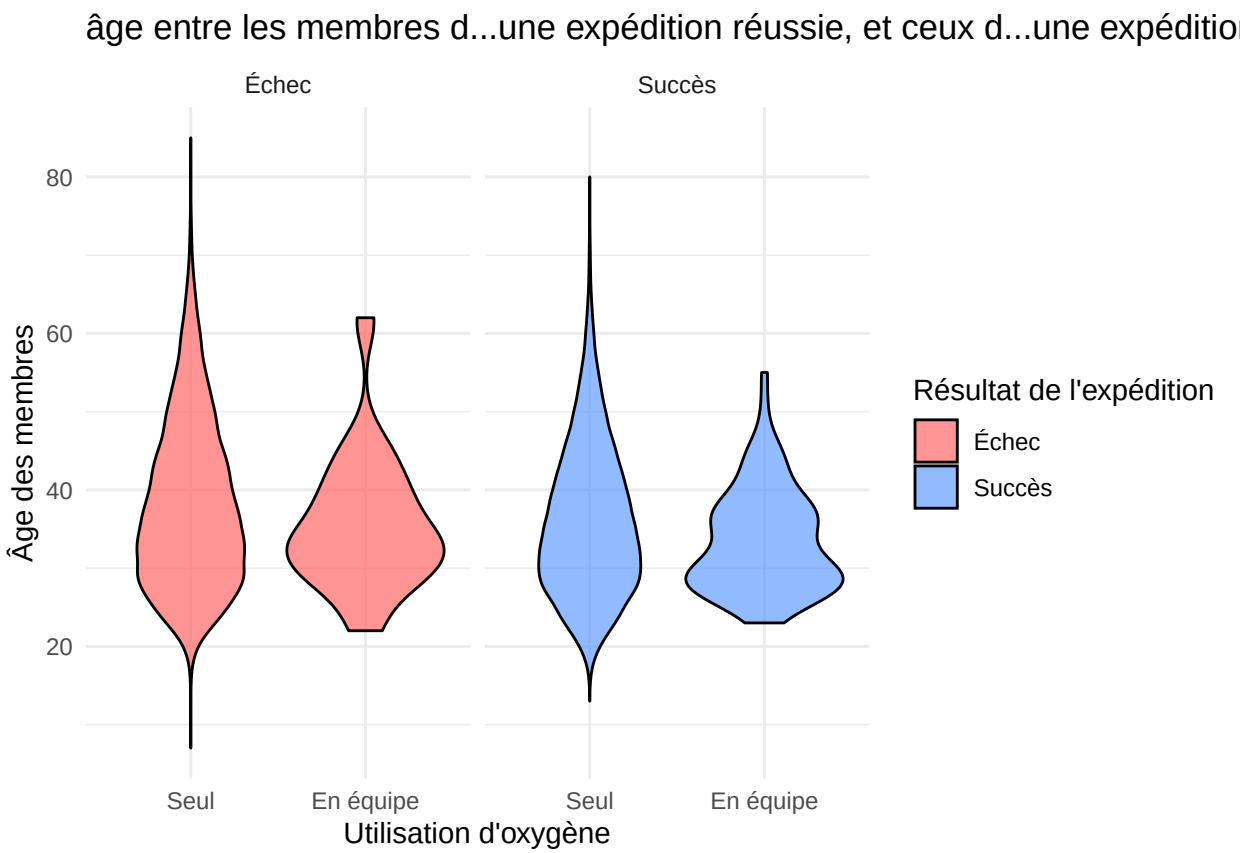
Échantillon: l'ensemble des alpiniste dont l'age est recensé dans ce csv

Variables : Le statut de l'expédition (réussie ou échouée), la présence d'une équipe, l'âge des membres.

Modalités : Pour le statut de l'expédition, la modalité est soit "réussie" soit "échouée". Pour la réalisation en équipe, la modalité est soit "En équipe" soit "Seul". Pour l'âge des membres, la modalité est un nombre.

```
tab <- subset(expeditions, !is.na(age))

ggplot(tab, aes(x = solo, y = age, fill = success)) +
  geom_violin(alpha = 0.7, color = "black") +
  facet_wrap(success ~ ., labeller = as_labeller(c(`TRUE` = "Succès", `FALSE` = "Échec"))) +
  labs(x = "Utilisation d'oxygène", y = "Âge des membres", fill = "Résultat de l'expédition") +
  ggtitle("âge entre les membres d'une expédition réussie, et ceux d'une expédition qui a échoué, avec
  scale_x_discrete(labels = c("Seul", "En équipe")) +
  scale_fill_manual(values = c("#FF6666", "#619CFF"), labels = c("Échec", "Succès")) +
  theme_minimal()
```



On remarque ici que les personnes faisant des expéditions seules ont tendance à avoir des âges beaucoup plus étalés, que l'expédition réussisse ou non. Pour ce qui est des expéditions en groupe l'âge est plus serré avec un écart entre 22 et 52 ans.