
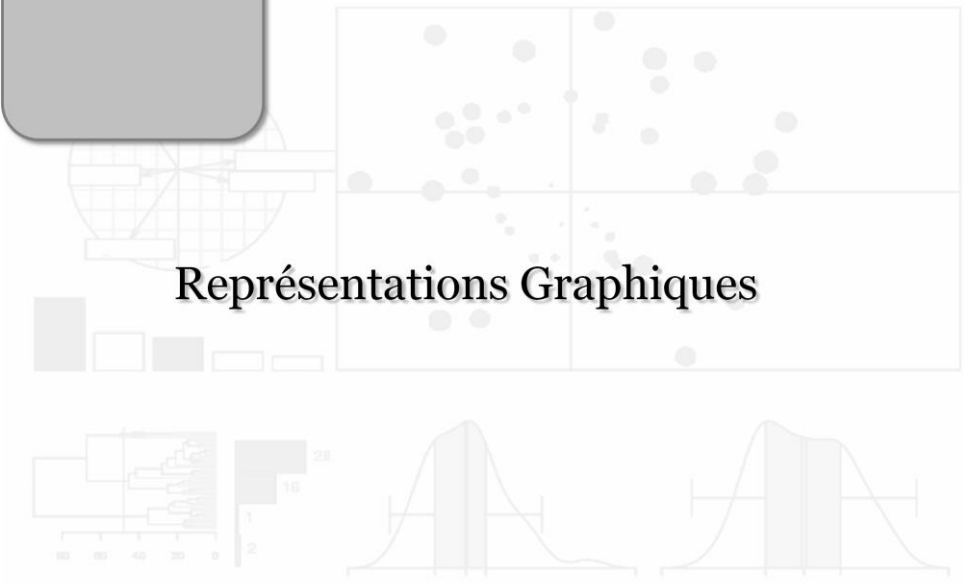




Chapitre 2
Introduction à la statistique avec R






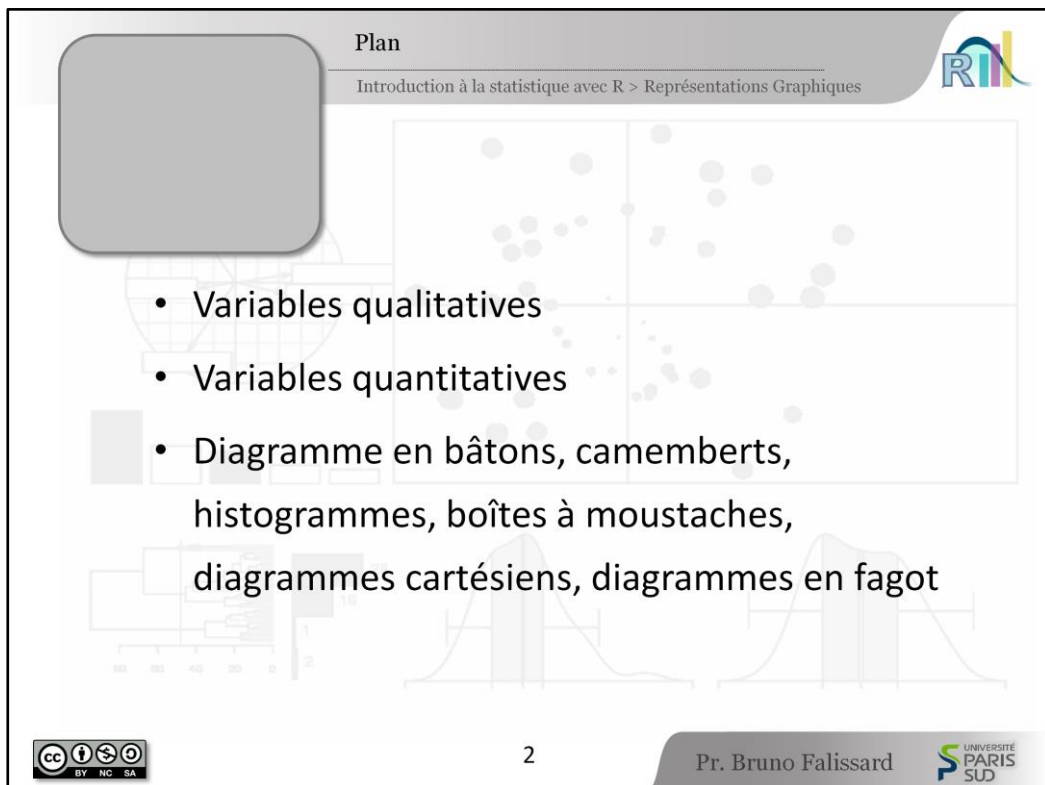
Représentations Graphiques


1
Pr. Bruno Falissard



[0:01] Pour pas mal d'utilisateurs des statistiques, plus l'outil statistique est compliqué, plus ils sont forts et plus ils vont être susceptibles de tirer la substance moelle de leur jeu de données. En réalité, l'expérience prouve que c'est à peu près le contraire. Plus une méthode statistique est simple, plus elle est efficace parce que tout le monde comprend les résultats. Le plus simple finalement, c'est les représentations graphiques et c'est pour ça que ce deuxième cours est particulièrement important. Le seul inconvénient des méthodes graphiques, c'est qu'elles prennent beaucoup de place.

Plan

Introduction à la statistique avec R > Représentations Graphiques




- Variables qualitatives
- Variables quantitatives
- Diagramme en bâtons, camemberts, histogrammes, boîtes à moustaches, diagrammes cartésiens, diagrammes en fagot



2


Pr. Bruno Falissard

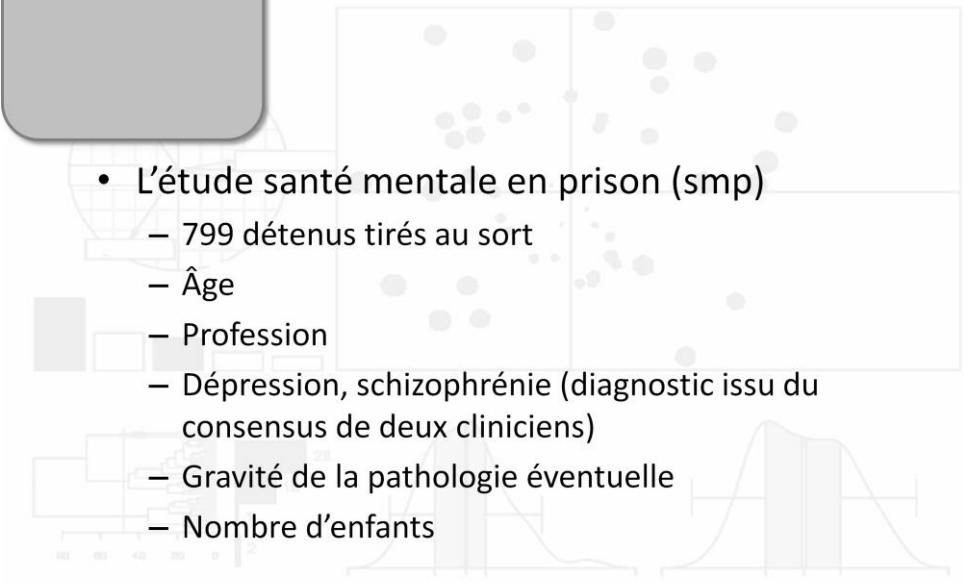


[0:33] Nous verrons successivement dans ce cours comment représenter graphiquement la distribution de variables qualitatives puis de variables quantitatives. Nous verrons plus précisément les diagrammes en bâtons, les camemberts, les histogrammes, les boîtes à moustaches, les diagrammes cartésiens et les diagrammes en fagot.


Le fichier smp.c

Introduction à la statistique avec R > Représentations Graphiques






- L'étude santé mentale en prison (smp)
 - 799 détenus tirés au sort
 - Âge
 - Profession
 - Dépression, schizophrénie (diagnostic issu du consensus de deux cliniciens)
 - Gravité de la pathologie éventuelle
 - Nombre d'enfants



3

Pr. Bruno Falissard



[0:53] Pour commencer, faisons connaissance avec le fichier `smp` que nous allons utiliser dans le restant du cours.

Ce fichier est relatif à l'étude `santé mentale en prison`, réalisée en 2004 et financée par le Ministère de la Justice et le Ministère de la Santé. Cette étude a porté sur 799 détenus masculins tirés au sort dans les prisons de France métropolitaine. Nous avons ici un extrait de 9 variables avec :

- l'âge
- la profession du détenu
- l'existence d'un diagnostic de dépression, de schizophrénie, réalisé par 2 cliniciens, c'est un diagnostic consensuel
- le niveau de gravité éventuelle de la pathologie du détenu, ici gravité consensuelle également
- le nombre d'enfants du détenu

- L'étude santé mentale en prison (smp), variables évaluant la personnalité des détenus
 - Recherche de sensation (rs) : curiosité, attrait pour le risque et la nouveauté
 - Évitement du danger (ed) : timidité, précautionneux
 - Dépendance à la récompense (dr) : sensibilité aux relations sociales, influençable

[1:40] et puis 3 variables relatives à sa personnalité :

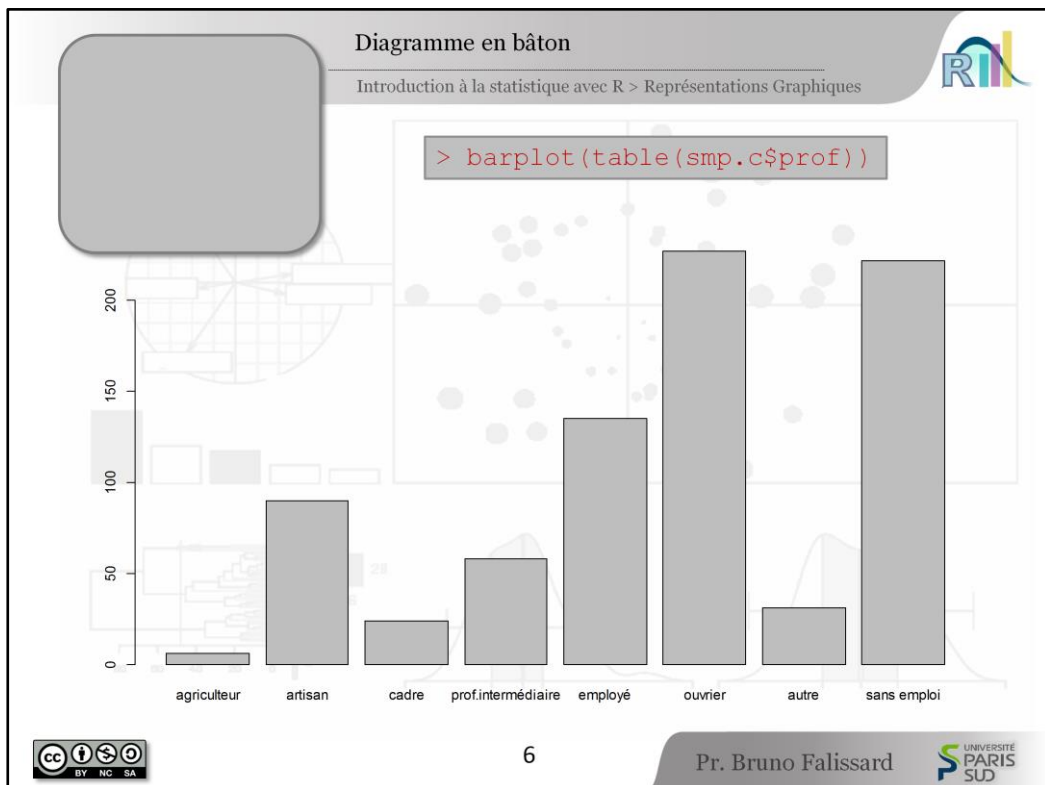
- le niveau de recherche de sensation (rs)
- le niveau d'évitement du danger (ed)
- et le niveau de dépendance à la récompense (dr)

```
> smp.c <- read.csv2("D:/MOOC/Data/smpl.csv")
> str(smp.c)
'data.frame': 799 obs. of 9 variables:
 $ age      : int  31 49 50 47 23 34 24 52 42 45 ...
 $ prof     : Factor w/ 8 levels "agriculteur",...: 3 NA 7 6...
 $ dep.cons : int   0 0 0 0 1 0 1 0 1 0 ...
 $ scz.cons : int   0 0 0 0 0 0 0 0 0 0 ...
 $ grav.cons: int   1 2 2 1 2 1 5 1 5 5 ...
 $ n.enfant : int   2 7 2 0 1 3 5 2 1 2 ...
 $ rs       : int   2 2 2 2 2 1 3 2 3 2 ...
 $ ed       : int   1 2 3 2 2 2 3 2 3 2 ...
 $ dr       : int   1 1 2 2 2 1 2 2 1 2 ...
```

[1:55] Vous importez le fichier le fichier `csv` à l'aide de l'instruction `read.csv2()`.

La chose à faire immédiatement après est naturellement de vérifier le contenu du fichier qu'on a importé.

La solution la plus simple est sûrement d'utiliser l'instruction `str(le-nom-du-fichier)` qui vous décrit son contenu, le nombre de sujets, le nombre de variables, l'intitulé des variables et contenu des premières valeurs.



[2:20] Une façon classique de représenter la distribution d'une variable aléatoire qualitative, c'est d'utiliser un diagramme en bâtons.

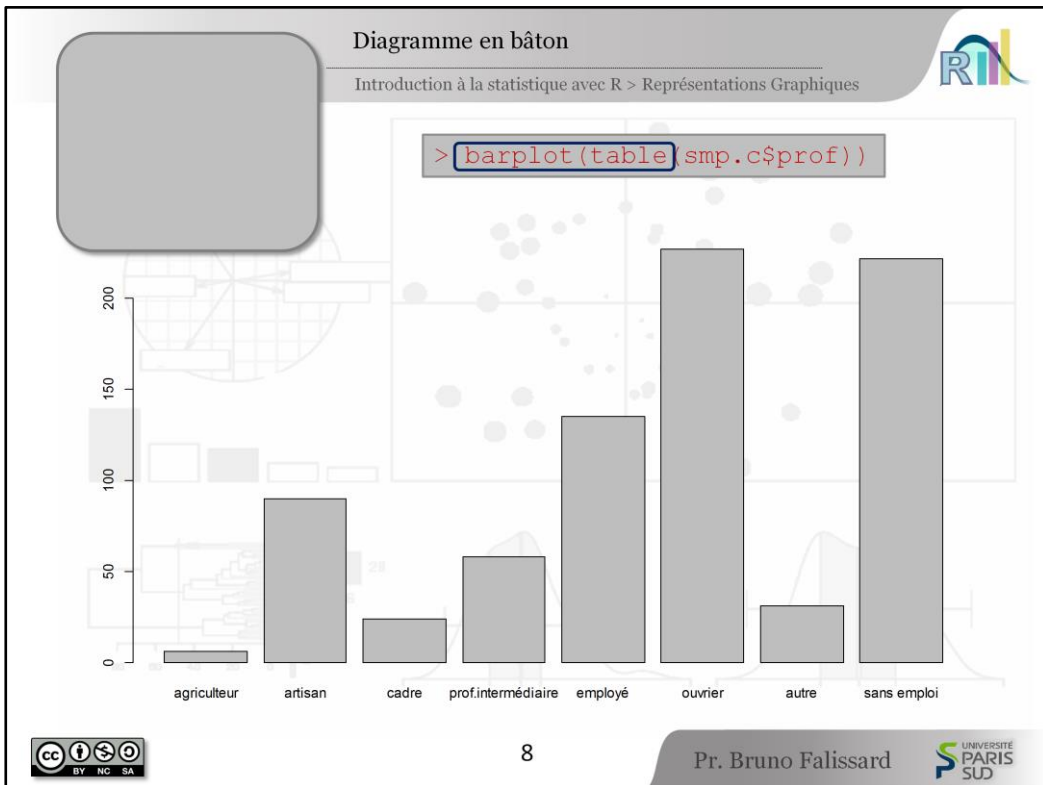
Avec R, il faut utiliser les fonctions `barplot()` et `table()`.

```
> barplot(table(smp.c$prof))
```

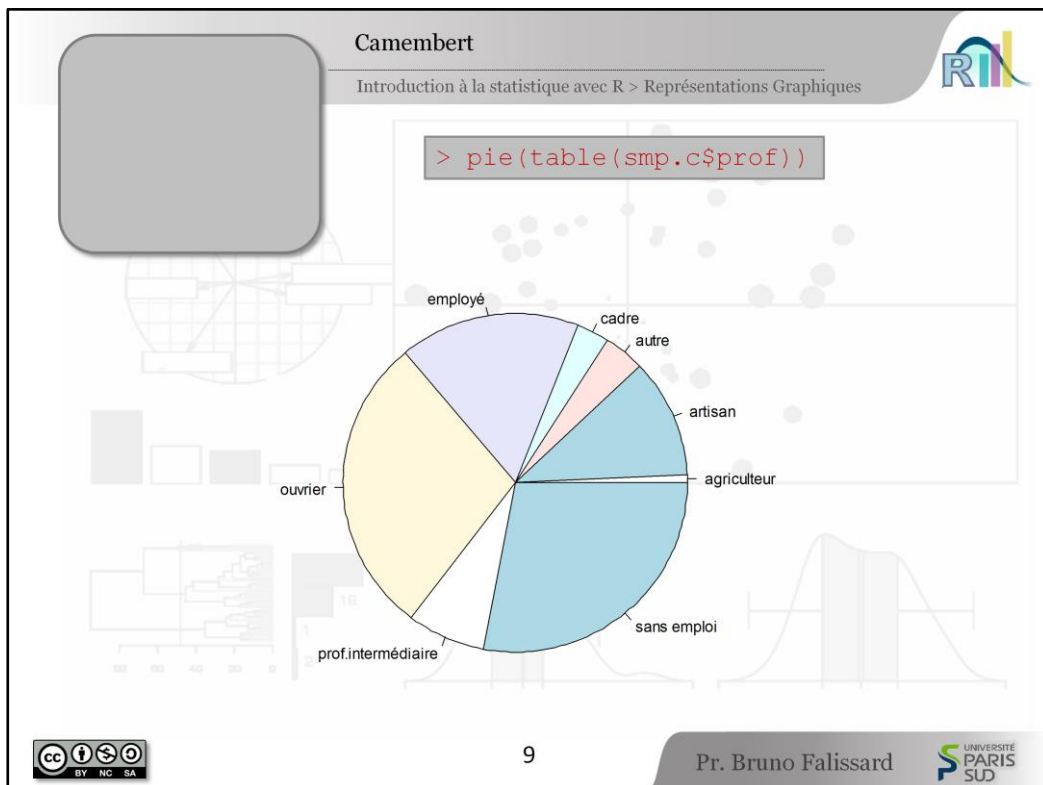
```
> str(smp.c$prof)
Factor w/ 8 levels "agriculteur",...: 3 NA 7 6 8 6 3 2 6 6 ...
> table(smp.c$prof)
```

agriculteur	artisan	autre
6	90	31
cadre	employé	ouvrier
24	135	227
prof.intermédiaire	sans emploi	
58	222	

[2:33] `table()` va calculer le nombre de détenus ayant chacun des métiers

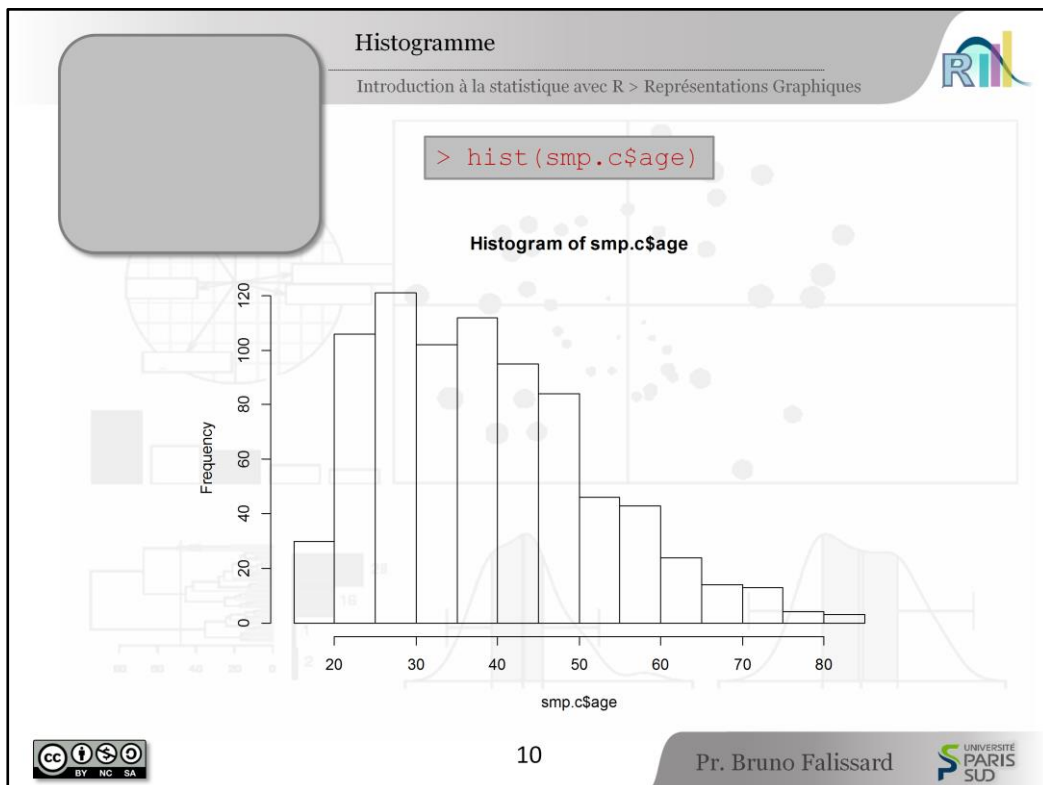


[2:37] et `barplot()` va représenter des bâtons ayant comme hauteur le nombre de ces détenus.



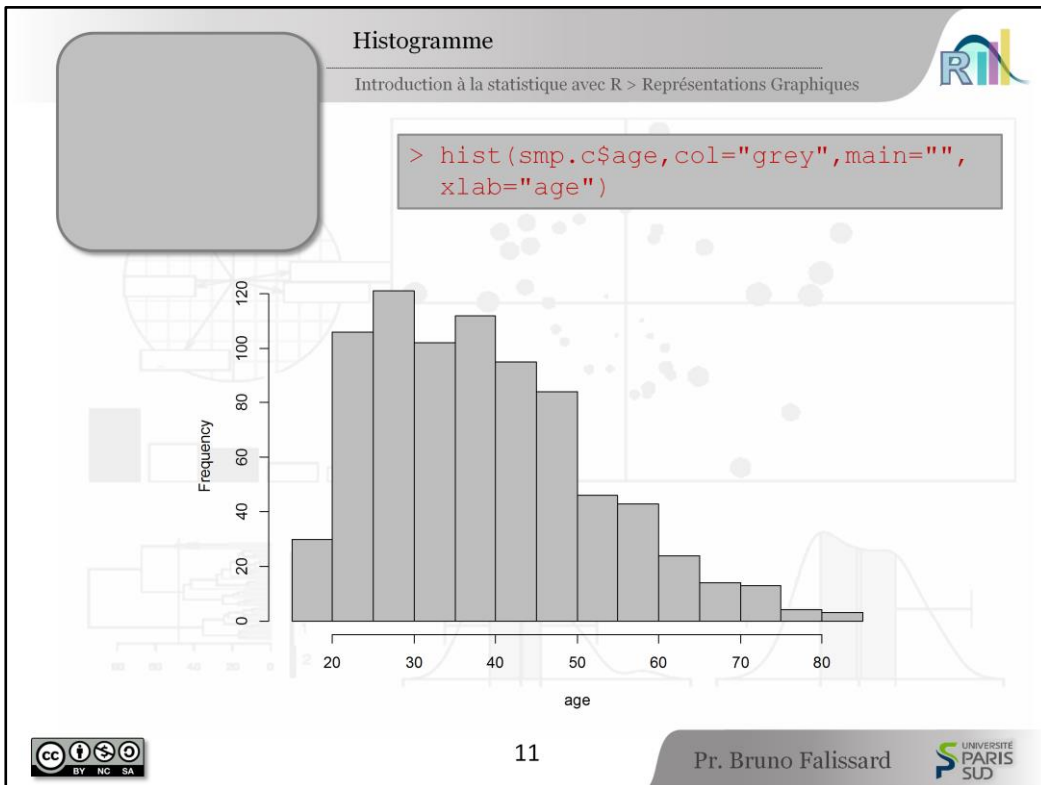
[2:44] Un autre grand classique pour représenter graphiquement la distribution d'une variable aléatoire catégorielle, c'est d'utiliser un camembert.

Avec R, on utilise les fonctions `pie()` et `table()`, "pie" signifiant en anglais "tarte". Certains statisticiens sont réticents à l'usage de ces camemberts : en effet, il semblerait que l'œil humain ait des difficultés à percevoir intuitivement des rapports de surface entre des secteurs de cercle, c'est-à-dire entre des parts de tarte, ou des parts de camembert. Alors qu'au contraire, l'œil humain serait capable de percevoir intuitivement des différences de hauteur de bâtons dans un diagramme en bâtons. En pratique, les représentations en camembert ont une certaine utilité quand on est intéressé par la part que représente une profession donnée par rapport à l'ensemble des détenus. En effet, chaque secteur peut être comparé à la superficie totale du cercle, ou du disque. Au contraire, avec un diagramme en bâtons, il faudrait avoir un bâton qui corresponde à l'ensemble de l'effectif étudié, ce qui serait peu commode.

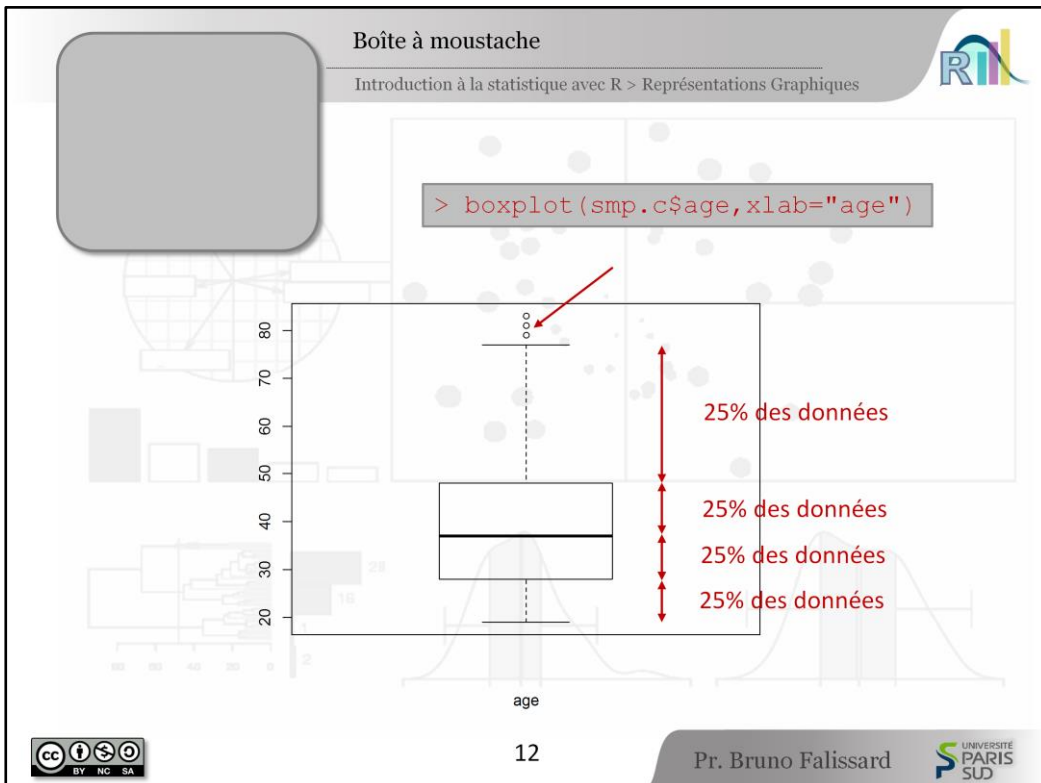


[4:00] Le grand classique pour représenter la distribution d'une variable aléatoire quantitative continue, c'est l'histogramme. Pour une variable aléatoire quantitative discrète, il vaut mieux utiliser un diagramme en bâtons. La différence entre les deux, c'est qu'avec l'histogramme, les bâtons sont contigus pour bien montrer qu'il y a une continuité dans les valeurs de la variable. Le seul point théorique un peu délicat avec un histogramme, c'est comment déterminer le nombre de bâtons. En pratique avec R, c'est automatique et 99 fois sur 100 ça marche très bien. L'instruction est très simple, c'est la fonction `hist(variable)`.

On peut être un peu déçu de l'aspect graphique notamment du fait que les bâtons ne sont pas grisés.



[4:42] Pour cela, il est possible d'ajouter des instructions à la fonction `hist()`, par exemple `col="grey"` pour avoir des bâtons grisés, et puis on peut décider de changer le titre du graphique et de changer la légende de l'axe des `x` comme ici : on a supprimé le titre du graphique avec `main=""` donc il n'y a aucun titre, et puis l'instruction `xlab` permet de déterminer la légende de l'axe des `x`.



[5:08] Une autre façon plus synthétique de représenter graphiquement la distribution d'une variable aléatoire quantitative, c'est d'utiliser une boîte à moustaches. L'instruction R est très simple, c'est la fonction `boxplot()`. Il suffit juste après de signifier la variable, ici l'âge du fichier `smp.c` et puis j'ai rajouté `xlab` pour bien indiquer en légende qu'il s'agit de la variable `age`.

Une boîte à moustaches s'interprète de la façon suivante :

- à l'intérieur de la boîte, vous avez 50 % des données,
- vous avez ensuite une moustache supérieure,
- et entre le bord supérieur de la boîte et la moustache supérieure, vous avez 25 % des données,
- et entre le bord inférieur de la boîte et la moustache inférieure, vous avez aussi 25% des données.

Alors ce que je dis en réalité est un peu faux.

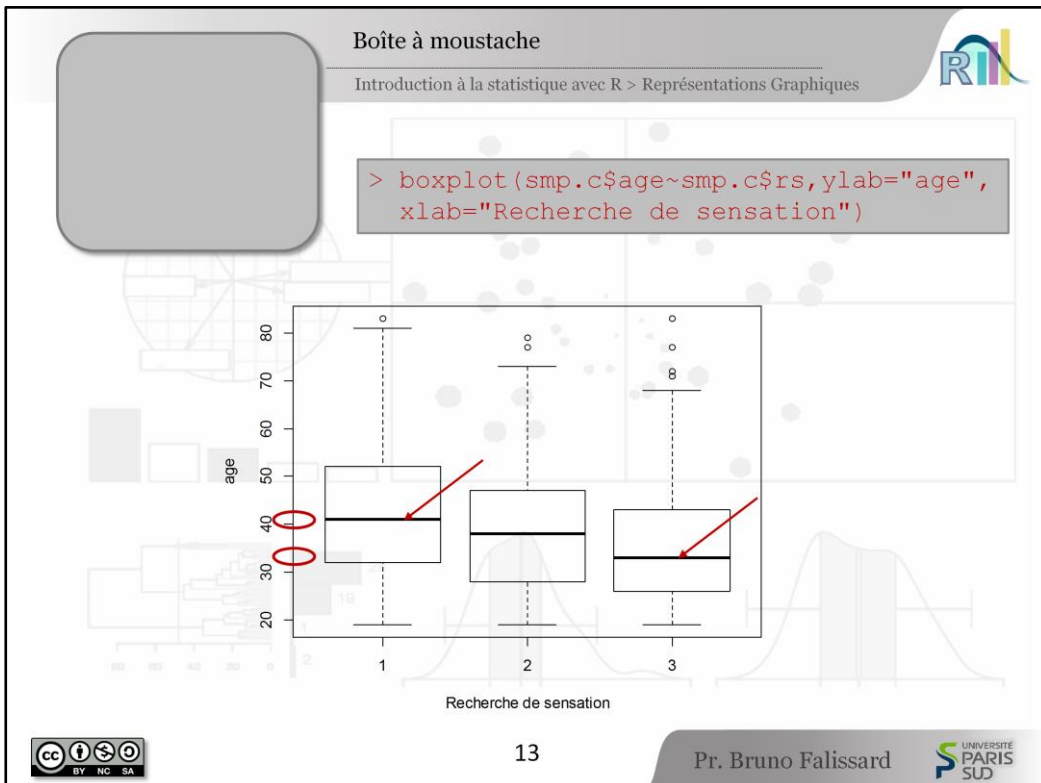
Si c'était vrai,

- la moustache supérieure devrait correspondre au maximum des données
- et la moustache inférieure au minimum
- et pourtant vous voyez sur ce graphique, au niveau de la flèche, quelques valeurs extrêmes que l'on appelle souvent "outliers".

En réalité, la définition de la moustache supérieure d'une boîte à moustaches est horriblement compliquée.

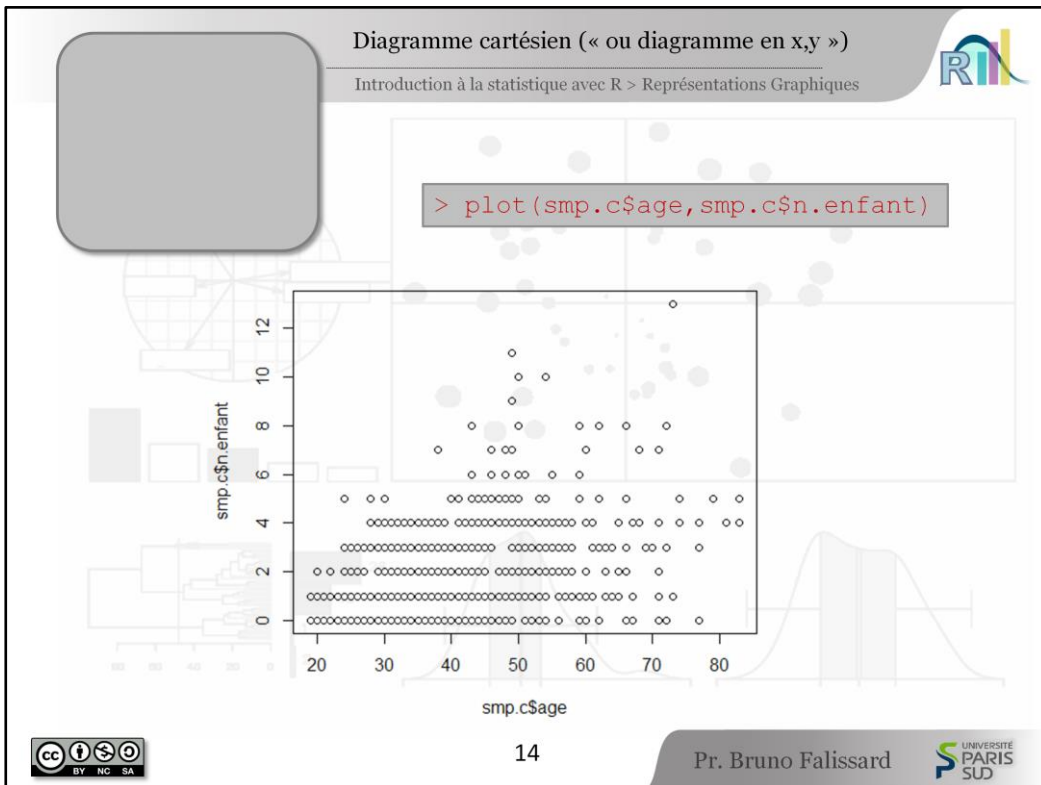
C'est le $\min(\max \text{ des données}, 1,5 \text{ écart-types au dessus du bord supérieur de la boîte})$.

C'est complètement incompréhensible et personne ne le retient.



[6:33] Les boîtes à moustaches sont réellement utiles pour représenter graphiquement la distribution d'une variable quantitative en fonction de sous-groupes. Par exemple, on pourrait se demander : "Est-ce que la distribution de l'âge est la même selon qu'on ait un niveau de recherche de sensations "faible", "moyen" ou "élevé" ?". C'est ce que nous avons fait sur le graphique présent.

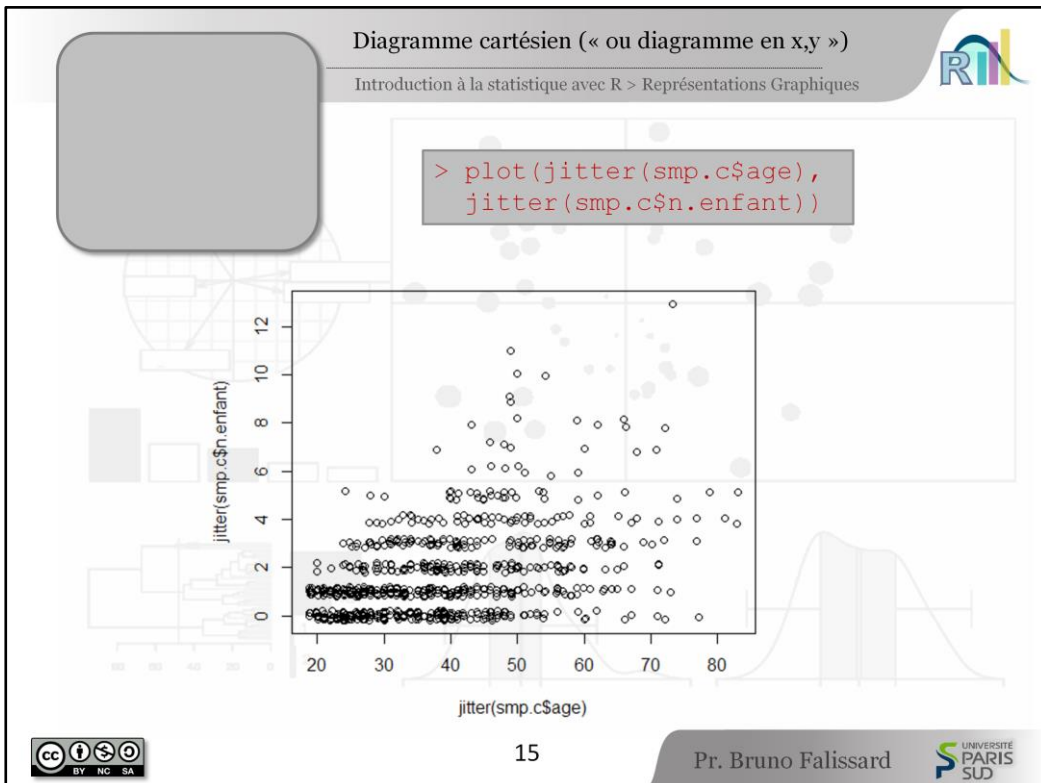
L'instruction R est aussi simple que précédemment, c'est `boxplot()` et il suffit à côté de la variable `age` de mettre un tilde (`~`) que l'on obtient à partir des touches `Alt Gr 2` et puis (de) la variable qui détermine les sous-groupes que l'on veut représenter, ici la variable `recherche de sensation`. On observe ici que, globalement, la (distribution est légèrement supérieure) distribution en âge est légèrement supérieure quand on a un faible niveau de sensation, plutôt que quand on a un niveau de sensation élevé.



[7:32] Pour représenter graphiquement la distribution conjointe de deux variables aléatoires quantitatives à l'aide d'un traditionnel graphique en x/y ou graphique cartésien, il faut utiliser la fonction `plot()` avec d'une part la variable qu'il y aura selon l'axe des x , et d'autre part la variable qu'il y aura selon l'axe des y .

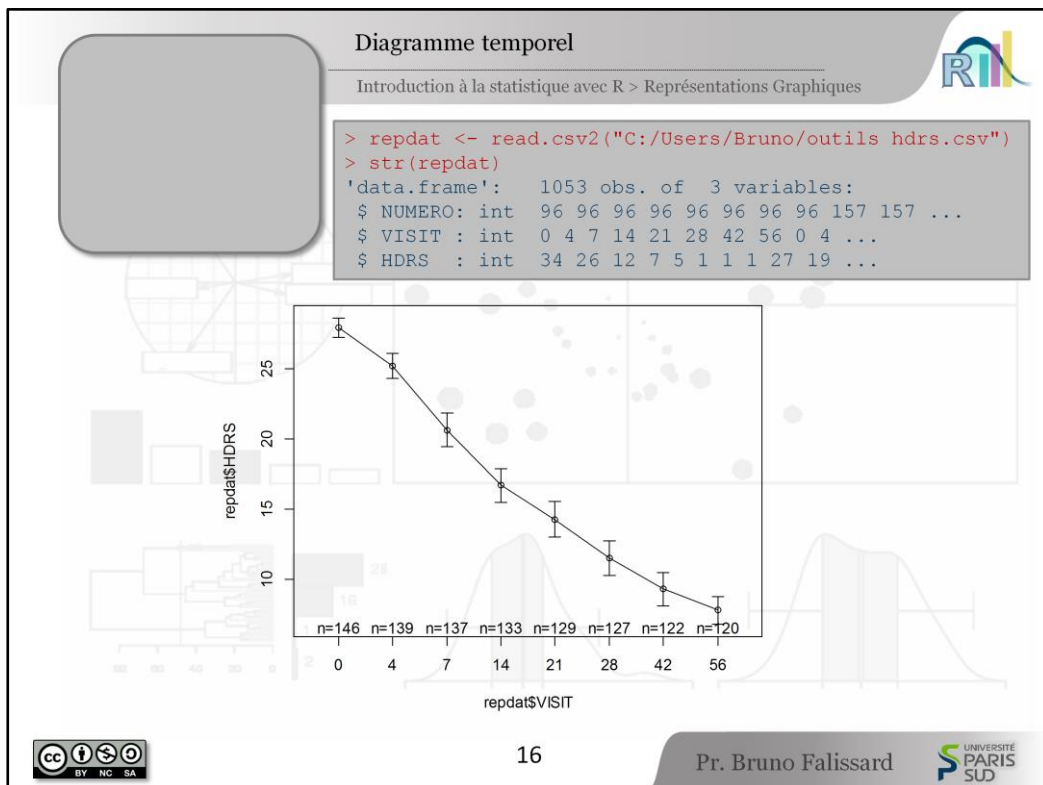
Dans le cas présent, nous avons représenté la distribution de l'âge et du nombre d'enfants et fort logiquement, plus un détenu est âgé, plus il a en moyenne un nombre d'enfants élevé.

On peut être surpris sur ce graphique par le fait qu'il semble ne pas y avoir 800 points correspondant aux 799 détenus, et cela s'explique naturellement : c'est que 2 détenus ayant chacun 50 ans et 2 enfants auront un point situé exactement au même endroit. Ça n'est pas gênant en soi, mais ça peut induire un peu en erreur. On peut avoir l'impression qu'il y a moins de sujets qu'il y en a réellement.



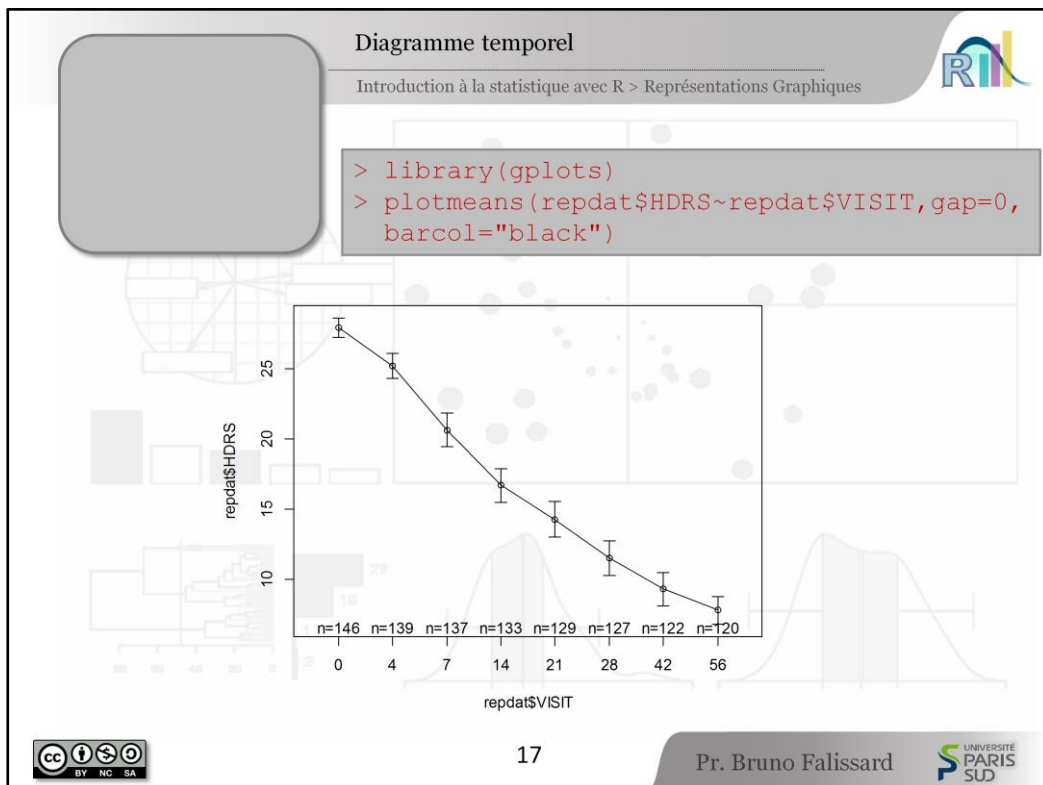
[8:32] Une façon de se tirer de ce faux pas, c'est de bouger légèrement de façon aléatoire chaque point de façon à ce qu'ils se décollent les uns des autres. L'instruction correspondante est la fonction `jitter()`.

Nous voyons ici `plot(jitter(age), jitter(n.enfant))` et nous avons un graphique plus agréable à regarder où cette fois, il y a bien 799 points.



[8:56] Parfois, c'est l'évolution temporelle de la distribution d'une variable aléatoire quantitative que l'on veut représenter. Le diagramme correspondant s'appelle diagramme temporel voire parfois diagramme de température.

Alors nous n'allons pas pouvoir utiliser le fichier `santé mentale en prison` parce que c'est une étude transversale en un temps donné donc on ne peut pas représenter graphiquement d'évolution au cours du temps. Pour cela, exceptionnellement nous allons utiliser un autre fichier de données. Ce sont des patients déprimés, hospitalisés pour dépression et qui sont traités et suivis pendant quelques semaines. L'instruction qui permet de représenter graphiquement l'évolution du score de dépression – ici c'est le score `hdrs` pour "hamilton depressive rating scale" – ...



[9:50] ... cette instruction, c'est la fonction `plotmeans()`. La fonction `plotmeans()` ne fait pas partie du bagage standard de R, elle fait partie de la librairie `gplots`.

Pour pouvoir l'utiliser, il faut d'abord que vous installiez le package `gplots`. Pour cela, une fois que vous avez ouvert R, vous devez aller dans le menu `packages` et vous cliquez sur `Installer les packages`, et là vous choisirez le site miroir de R, puis vous cliquerez sur `gplots` et elle sera téléchargée. Vous n'avez besoin de faire ça qu'une seule fois.

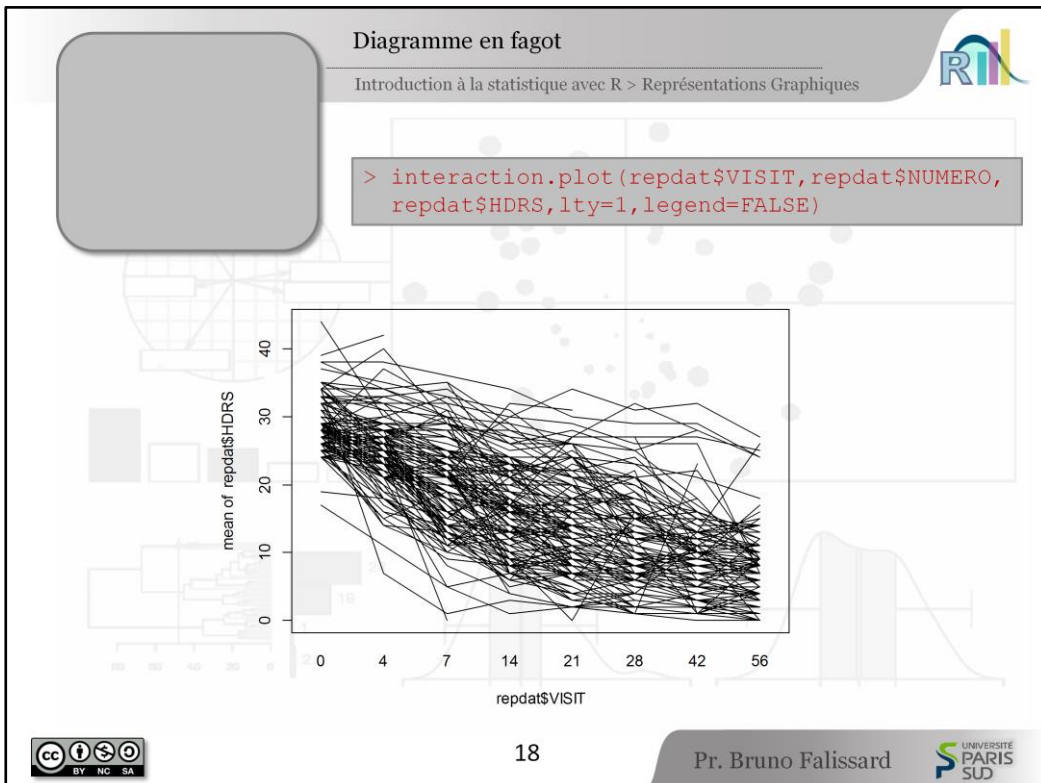
Donc ici nous appelons la librairie `gplots` et puis après nous appelons la fonction `plotmeans()` avec tout simplement

- la variable à représenter, ici la variable `HDRS`,
- un tilde (`~`),
- et puis la variable qui représente le temps, ici la variable `VISIT`.

Les instructions `gap` et `barcol` ne sont là que pour faire que la représentation graphique soit plus agréable à regarder. Nous constatons sur le dessin que progressivement, au fil du temps, l'état symptomatique des patients s'améliore progressivement.

NB : La fonction `plotmeans()` a évolué depuis la première session du MOOC et la commande

`plotmeans(repdat$HDRS~repdat$VISIT, gap=0, barcol="black")` produit un warning "`gap`" is not a graphical parameter mais les résultats sont bien ceux escomptés.



[10:34] Plutôt que de représenter l'évolution moyenne des sujets au cours du temps, il peut être intéressant de représenter l'évolution de chaque sujet. Bien sûr, avec plusieurs centaines d'individus dans un jeu de données, l'ensemble peut faire un peu fouillis. Néanmoins, cela donne une impression intéressante de la variabilité de l'évolution d'un sujet à l'autre.


La fonction correspondante est la fonction `interaction.plot()`. Sa syntaxe est très simple :

- vous utilisez d'abord comme variable la variable temporelle, ici la variable `VISIT`,
- puis la variable qui indique chaque sujet, ici la variable `NUMERO`,
- puis la variable que vous voulez représenter, ici le score de dépression `HDRS`.

Les instructions `lty=1` correspondent au fait que nous voulons des lignes droites continues et `legend` indique la légende.

Conclusion

Introduction à la statistique avec R > Représentations Graphiques



```

smp.c <- read.csv2("D:/MOOC/Data/smpl.csv")
str(smp.c)

barplot(table(smp.c$prof))
pie(table(smp.c$prof))


hist(smp.c$age)
hist(smp.c$age, col="grey", main="", xlab="age")
boxplot(smp.c$age, xlab="age")
boxplot(smp.c$age~smp.c$rs, ylab="age", xlab="Recherche de sensation")

plot(smp.c$age, smp.c$n.enfant)
plot(jitter(smp.c$age), jitter(smp.c$n.enfant))


repdat <- read.csv2("D:/MOOC/Data/hdrs.csv")
str(repdat)

library(gplots)
plotmeans(repdat$HDRS~repdat$VISIT, gap=0, barcol="black")
interaction.plot(repdat$VISIT, repdat$NUMERO, repdat$HDRS, lty=1, legend=FALSE)

```



19

Pr. Bruno Falissard


[11:47] A l'issue de ce cours,

- vous savez maintenant représenter graphiquement la distribution d'une variable qualitative à l'aide de diagrammes en bâtons avec la fonction `barplot()`, à l'aide de camemberts à l'aide de la fonction `pie()`,
- vous savez représenter une variable quantitative avec des histogrammes et la fonction `hist()` et des boîtes à moustaches avec la fonction `boxplot()`,
- vous savez représenter de façon conjointe la distribution de 2 variables aléatoires avec les diagrammes x/y avec la fonction `plot()`,
- et puis finalement, vous savez représenter l'évolution d'une variable au cours du temps à l'aide de la fonction `plotmeans()` ou de la fonction `interaction.plot()`.

Ce que je vous propose maintenant, c'est de faire une pause, d'ouvrir votre ordinateur, de lancer R et de refaire tourner toutes ces syntaxes pour retrouver les mêmes résultats.

Bon courage.