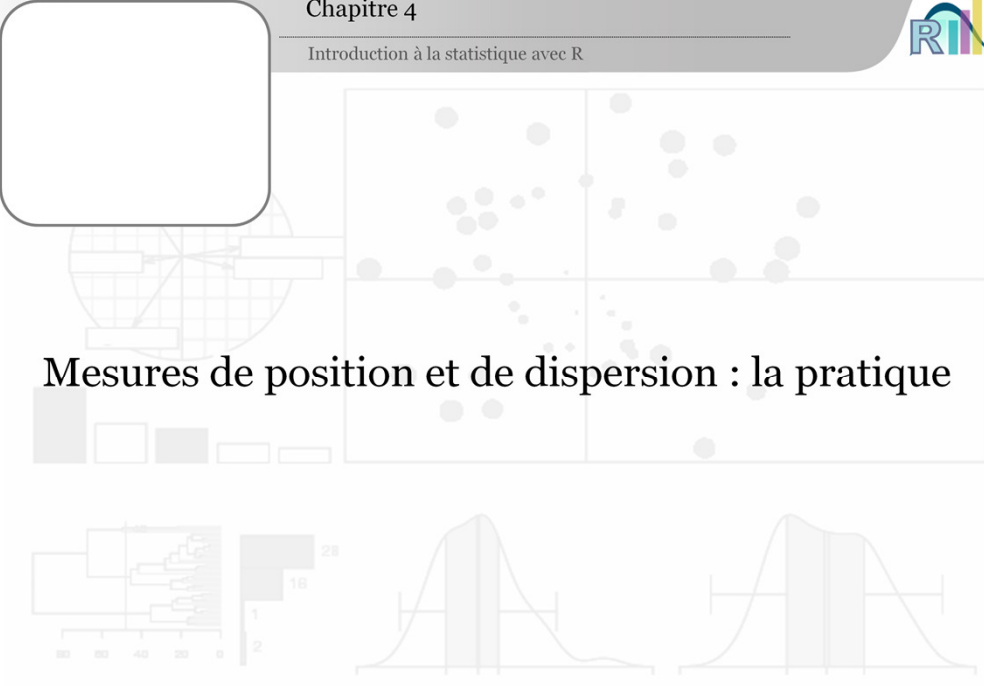



Chapitre 4

Introduction à la statistique avec R




Mesures de position et de dispersion : la pratique



1

Pr. Bruno Falissard



[0:01] Alors voyons maintenant comment calculer tous ces paramètres en pratique.

summary

Introduction à la statistique avec R > Position, dispersion : la pratique

R Console

```
> summary(smp.c)
```

<p>age</p> <p>Min. :19.0</p> <p>1st Qu.:28.0</p> <p>Median :37.0</p> <p>Mean :38.9</p> <p>3rd Qu.:48.0</p> <p>Max. :83.0</p> <p>NA's :2</p>	<p>ouvrier</p> <p>employé</p> <p>artisan</p> <p>prof.intermédiaire</p> <p>autre</p> <p>(Other)</p> <p>NA's</p>	<p>prof</p> <p>:227</p> <p>:135</p> <p>: 90</p> <p>: 58</p> <p>: 31</p> <p>: 30</p> <p>:228</p>	<p>dep.cons</p> <p>Min. :0.0000</p> <p>1st Qu.:0.0000</p> <p>Median :0.0000</p> <p>Mean :0.3967</p> <p>3rd Qu.:1.0000</p> <p>Max. :1.0000</p> <p>NA's</p>	<p>scz.cons</p> <p>Min. :0.0000</p> <p>1st Qu.:0.0000</p> <p>Median :0.0000</p> <p>Mean :0.0826</p> <p>3rd Qu.:0.0000</p> <p>Max. :1.0000</p> <p>NA's</p>
---	--	---	---	---

```
>
```

2

Pr. Bruno Falissard

[0:04] C'est très simple, il suffit d'utiliser la fonction `summary()` avec le nom de votre fichier, pour nous `smp.c`. Et puis le logiciel automatiquement,

d'une part pour les **variables quantitatives** vous propose

- le minimum,
- le 1^{er} quartile,
- la médiane,
- la moyenne,
- le 3^{ème} quartile,
- le maximum
- et le nombre de données manquantes, c'est très important de connaître le nombre de données manquantes pour chaque variable.


et puis pour les **variables catégorielles** comme profession vous avez

- le nombre de sujets pour chaque modalité
- ainsi que les données manquantes.

Donc cette fonction `summary()` est extrêmement utile. Elle a pour certains des inconvénients. Le premier, c'est qu'elle utilise beaucoup de place. Vous voyez pour chaque variable est répété minimum, médiane, etc. Et donc si vous avez plusieurs centaines de variables, à la fin ce n'est pas très synthétique. Il est plus traditionnel de présenter les résultats avec les variables en ligne, et en colonnes : moyenne, médiane, quartile, etc. C'est plus facile à lire. Donc, des développeurs ont proposé une autre fonction qui est la fonction `describe()`.

describe

Introduction à la statistique avec R > Position, dispersion : la pratique




Fichier Edition Voir Misc Packages Fenêtres Aide

R Console

```
> library(prettyR)
> describe(smp.c)
Description of smp.c


Numeric
      mean  median    var    sd  valid.n
age      38.9     37  176.4  13.28     797
dep.cons 0.3967     0  0.2396  0.4895     799
scz.cons 0.0826     0  0.07587  0.2755     799
grav.cons 3.643     4   2.726  1.651     795
n.enfant 1.755     1   3.364  1.834     773
rs        2.057     2  0.7708  0.8779     696
ed        1.866     2   0.759  0.8712     692
dr        2.153     2   0.6885  0.8297     688

Factor
prof
Value Count Percent
NA      228  28.54
ouvrier 227  28.41
employé 135  16.9
artisan 90  11.26
prof.intermédiaire 58  7.26
autre    31  3.88
cadre    24  3
agriculteur 6  0.75
mode=ouvrier Valid n=571
> |
```



3

Pr. Bruno Falissard



[1:15] Elle n'est pas disponible en standard dans R. Il faut utiliser le package `prettyR`. Vous vous souvenez, il faut cliquer sur le menu `package/installer le package` et puis vous faites ça une fois seulement et après vous appelez `prettyR` par `library(prettyR)` et la fonction `describe()` de votre fichier. Et là, c'est vrai que la présentation est plus élégante.


Vous avez d'un côté, toutes les variables numériques quantitatives avec en lignes le nom des variables et en colonnes : moyenne, médiane, variance, écart-type, nombre de sujets disponibles.

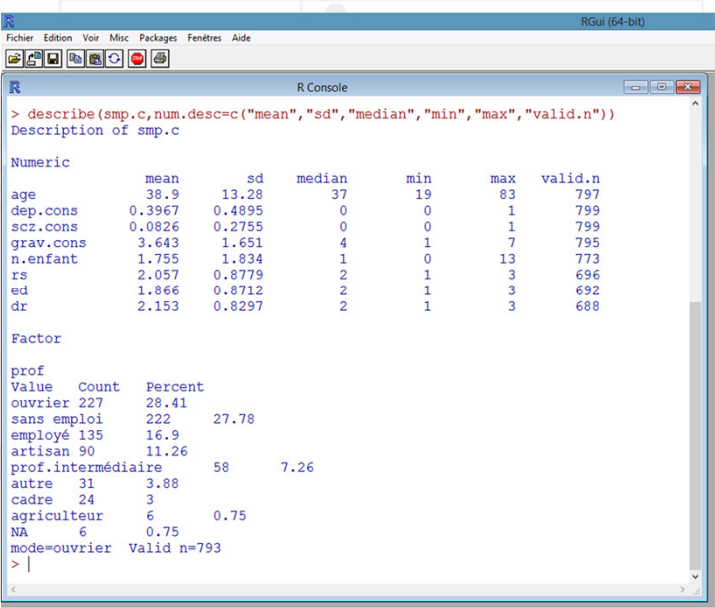
Et puis pour les variables catégorielles, à la suite, vous avez le nombre de modalités, avec le nombre de sujets et le pourcentage des sujets dans chaque modalité.

La fonction `describe()` est très intéressante. Elle a quand même un inconvénient, c'est qu'elle ne présente pas les quartiles. Ce n'est peut-être pas toujours indispensable mais elle ne présente ni le minimum, ni le maximum et ça, c'est absolument indispensable, parce que quand vous présentez un fichier, le minimum et le maximum vous permettent de détecter les données aberrantes. Par exemple, si vous avez la variable `age` et que vous avez une erreur de mesure ou une erreur de saisie et que vous avez un âge de 250 ans, s'il y a 1000 sujets vous risquez de passer à côté. Quand vous avez le maximum, ça sort automatiquement le maximum de l'âge, c'est 250 ans. Vous savez qu'il y a un problème. Donc `describe()` ne va pas, mais par chance, on peut demander à introduire des mesures supplémentaires.

describe

Introduction à la statistique avec R > Position, dispersion : la pratique






```


> describe(smp.c,num.desc=c("mean","sd","median","min","max","valid.n"))
Description of smp.c

Numeric
      mean      sd    median    min     max  valid.n
age      38.9    13.28       37     19      83      797
dep.cons 0.3967  0.4895        0      0       1      799
scz.cons 0.0826  0.2755        0      0       1      799
grav.cons 3.643  1.651         4      1       7      795
n.enfant  1.755  1.834         1      0      13      773
rs        2.057  0.8779        2      1       3      696
ed        1.866  0.8712        2      1       3      692
dr        2.153  0.8297        2      1       3      688

Factor
prof
Value Count Percent
ouvrier 227      28.41
sans emploi 222      27.78
employé 135      16.9
artisan 90      11.26
prof.intermédiaire 58      7.26
autre 31      3.88
cadre 24      3
agriculteur 6      0.75
NA 6      0.75
mode=ouvrier Valid n=793
  
```



4

Pr. Bruno Falissard
 


[2:49] C'est ce qui est fait ici : vous avez `describe()` toujours, le nom de votre fichier et puis vous demandez spécifiquement de calculer


- moyenne,
- écart-type,
- on a enlevé la variance parce que si on a l'écart-type, ce n'est pas tellement utile d'avoir la variance
- la médiane,
- le minimum,
- le maximum
- et le nombre de sujets sur lesquels on a la mesure.
- On aurait pu rajouter les quartiles aussi.

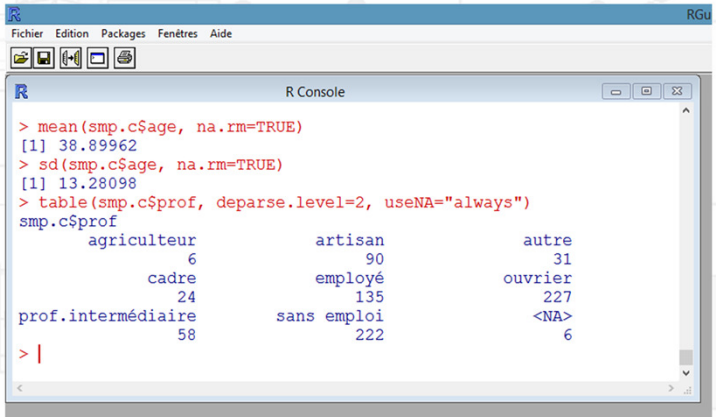
Alors pour terminer, il est parfois utile de calculer juste la moyenne, ou juste l'écart-type.

mean, sd, table

Introduction à la statistique avec R > Position, dispersion : la pratique









```

> mean(smp.c$sage, na.rm=TRUE)
[1] 38.89962
> sd(smp.c$sage, na.rm=TRUE)
[1] 13.28098
> table(smp.c$prof, deparse.level=2, useNA="always")
smp.c$prof
  agriculteur      artisan      autre
           6          90          31
      cadre    employé    ouvrier
       24       135       227
prof.intermédiaire sans emploi    <NA>
       58       222           6
  
```



5


Pr. Bruno Falissard
 


[3:16] Pour ça, vous avez la fonction `mean()` ou la fonction `sd()`, `sd` pour "standard deviation" qui veut dire écart-type.

Et pour une variable catégorielle, si vous voulez connaître les modalités, vous pouvez utiliser la fonction `table()`. La fonction `table()` est ici utilisée avec les options `deparse.level=2`. Ça permet d'avoir le nom de la variable affiché et puis `useNA` qui est utilisé ici pour pouvoir savoir combien de sujets ont eu des données manquantes pour leurs professions.



Conclusion


Introduction à la statistique avec R > Position, dispersion : la pratique






```
summary(smp.c)
library(prettyR)
describe(smp.c)
describe(smp.c,num.desc=c("mean","sd","median","min","max","valid.n"))
mean(smp.c$age, na.rm=TRUE)
sd(smp.c$age, na.rm=TRUE)
table(smp.c$prof, deparse.level=2, useNA="always")
```


6

Pr. Bruno Falissard


[3:44] Nous avons donc calculé moyenne, médiane, quartiles, écart-type, minimum, maximum, etc. à l'aide des fonctions `summary()`, `describe()`, `mean()`, `sd()`, `table()`.

Je vous propose maintenant comme d'habitude de prendre votre propre ordinateur et de refaire les calculs.