Capsule 4: Notions de statistique

Jane Coulombe
2022-08-01

Moyenne

La moyenne théorique, aussi appelée espérance, est généralement désirée par ce. On suppose ici des variables que x_i, une variable aléatoire continue, comme le pression sanguin, provenant d'une distribution de moyenne théorique μ. En faisant cette hypothèse, on assume que la distribution des données a une certaine "forme" tel., cloche, pour une variable aléatoire de la Normale).

On souhaite ensuite inférer sur cette distribution, par exemple, estimer cette moyenne μ à partir des statistiques de la variable aléatoire (i.-à-d. les données). Pour ce faire, on estimator sans biais pour μ est la moyenne échantillonnale.

La moyenne échantillonnale est égale à la somme des observations divisée par leur nombre, ce nombre correspondant généralement à la taille échantillonnale si toutes les données sont observées et qu'il n'y a donc pas de données manquantes.

Soit une variable aléatoire X et n réalisations de cette variable (données par $x_1, ..., x_n$), on a

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

La moyenne ou moyenne échantillonnale, dénotée $\bar{x}$, représente une mesure de "centre" de la distribution de vos données.

Malheureusement, la moyenne est influencée par les données extrêmes dans votre échantillon. La graphique suivant nous montre la distribution empirique de 1000 réalisations d'une variable aléatoire de la Gamma fortement asymétrique (étudie ainsi la moyenne échantillonnale en bleu.

```
set.seed(1000)
n <- 1000
mean_gamma_estimator_sample1 <- 0
for(i in 1:n){
mean_gamma_estimator_sample1 <- c(mean_gamma_estimator_sample1, mean(rgamma(n,1,1)))
}
mean_gamma_estimator_sample1 <- mean_gamma_estimator_sample1[-1]
hist(mean_gamma_estimator_sample1,
     main = "Histogram of data",
     xlab = "data",
     ylab = "Frequency",
     col = "lightblue",
     abline(v = mean(mean_gamma_estimator_sample1), col="red"))
```

Ce que l'on note, c'est que la moyenne (en bleu) est à droite de la majorité des données (qui sont plutôt concentrées à gauche dans la distribution (o-haut).

Médiane

La médiane est une autre mesure de centre de la distribution qui, elle, est moins influencée par les valeurs extrêmes dans la distribution.

Elle peut être calculée comme suit. On doit mettre toutes les réalisations d'une variable (données) en ordre croissant et choisir la donnée qui se retrouve exactement au milieu de cette suite (si le nombre de données est impair) ou la moyenne des deux données du centre (si le nombre de données est pair).

Example

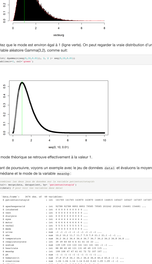
Soit X les 10 valeurs de la variable (qui suivent): 47, 21, 45, 90, 12, 3, 35, 82, 76, 75. On met d'abord les données en ordre croissant, ce qui mène à la suite:
3, 12, 21, 35, 45, 56, 47, 75, 82, 90. Comme la taille de l'échantillon est 10, les deux valeurs au centre sont celles placées en positions 5 et 6. Ainsi, la médiane est égale à
(45 + 56)/2 = 50.5 : Notez que, dans cet example, la médiane n'est pas commune dans les valeurs de l'échantillon!

En R, on peut le faire comme suit:

```
set.seed(1000)
sample1 <- c(47, 21, 45, 90, 12, 3, 35, 82, 76, 75)
sample1

## [1] 47 21 45 90 12 3 35 82 76 75
```

Ajoutons maintenant la médiane (en rouge) sur le même graphique vu plus tôt:

```
set.seed(1000)
n <- 1000
mean_gamma_estimator_sample1 <- 0
for(i in 1:n){
mean_gamma_estimator_sample1 <- c(mean_gamma_estimator_sample1, mean(rgamma(n,1,1)))
}
mean_gamma_estimator_sample1 <- mean_gamma_estimator_sample1[-1]
hist(mean_gamma_estimator_sample1,
     main = "Histogram of data",
     xlab = "data",
     ylab = "Frequency",
     col = "lightblue",
     abline(v = mean(mean_gamma_estimator_sample1), col="red"),
     abline(v = median(mean_gamma_estimator_sample1), col="blue"))

## Warning: Ignoring unknown parameters: col

## [1] 0.5
```



On remarque que la médiane est plus à gauche de la distribution que la moyenne. La médiane coupe l'ensemble des données en deux parties. Ainsi, comparée à la moyenne, elle n'est à peu près pas affectée par ce que se trouve aux extrémités du jeu de données.

Mode

Le mode (empirique) est la valeur de la variable avec la plus grande fréquence observée, c.-à-d. la valeur de la variable que l'on retrouve le plus souvent dans les données.

Par example, soient les données 1, 2, 3, 76, 23, 45, 2, 23, 56, 1, 1, 45.7, 2, 78, 23, 45, 43, 1, 2. La valeur le plus fréquente est le 1 qui retrouve aussi 4 fois dans nos données. On a un mode en, aussi de 1 et 2.

En R, il n'est pas de fonction toute prête en R qui permette de calculer le mode. Voici un example d'une fonction trouvée en ligne (au https://stackoverflow.com/questions/55/how-to-find-the-statistical-mode):

```
sample1 <- c(1, 2, 3, 76, 23, 45, 2, 23, 56, 1, 1, 45.7, 2, 78, 23, 45, 43, 1, 2)

Mode <- function() {
  # find the unique values
  ux <- unique(sample1)
  ux[which.max(tabulate(match(sample1, ux)))]
}

Mode()

## [1] 1
```

Vous pouvez aussi créer votre propre fonction pour calculer le mode, mais n'oubliez pas de considérer les cas où on a plusieurs modes.

Le mode théorique correspond au plus haut point ("bosse") de la distribution théorique. Par example, simulons plusieurs variables aléatoires de la Gamma(1, 3) et visualisons la distribution empirique de ces réalisations (qui devrait ressembler beaucoup à la distribution théorique).

```
set.seed(1000)
n <- 1000
hist(rgamma(n,1,3),main="Histogram of rectangular",
     xlab="data",
     ylab="Frequency",
     col="lightblue")
```

Notez que le mode est environ égal à 1 (ligne verte). On peut regarder la vraie distribution d'une variable aléatoire Gamma(3, 3) comme suit:

```
plot(dgamma(seq(0,3,0.01),3,3), ylab="dgamma(x,3,3)",
     xlab="x", type="l")
```

Le mode théorique se retrouve effectivement à la valeur 1.

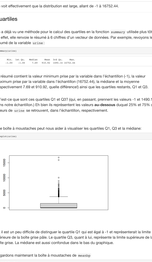
Avant de poursuivre, voyons un example avec le jeu de données `data()` et évaluons la moyenne, la médiane et le mode de la variable `mpg`:

```
# Calculer des deux mode et moyenne pour la variable mpg
# Utiliser le jeu de données mtcars
mtcars[order(-mpg),c(1:2,9:11,16:17)]

##                mpg cyl am gear carb
## Fiat 128      32.4   4  1    4    1
## Honda Civic   30.4   4  1    4    2
## Toyota Corolla 33.9   4  1    4    1
## Fiat X1-9     27.3   4  1    4    1
## Porsche 914-2 26.0   4  1    5    2
## Lotus Europa  30.4   4  1    5    2
## Ford Pantera L 15.8   8  1    5    4
## Ferrari Dino  19.7   6  1    5    6
## Maserati Bora 15.0   8  1    5    8
## Merc 280      19.2   6  0    4    4
## Merc 280C     17.8   6  0    4    4
## Merc 450SE    16.4   8  0    3    3
## Merc 450SL    17.3   8  0    3    3
## Merc 450SLC   15.2   8  0    3    3
## Cadillac Fleetwood 10.4   8  0    3    4
## Lincoln Continental 10.4   8  0    3    4
## Chrysler Imperial 14.7   8  0    3    4
## Dodge Challenger 15.5   8  0    3    2
## AMC Javelin   15.2   8  0    3    2
## Camaro Z28    13.3   8  0    3    4
## Pontiac Firebird 19.2   8  0    3    2
## Hornet 4 Drive 21.4   6  0    3    1
## Hornet Sportabout 18.7   8  0    3    2
## Valiant       18.1   6  0    3    1
## Merc 240D     24.4   4  0    4    2
## Merc 230      22.8   4  0    4    2
## Mazda RX4     21.0   6  1    4    4
## Mazda RX4 Wag 21.0   6  1    4    4
## Merc 240C     24.4   4  0    4    2
## Duster 360    14.3   8  0    3    4
## Merc 450SLC   15.2   8  0    3    3
## Cadillac Fleetwood 10.4   8  0    3    4
## Lincoln Continental 10.4   8  0    3    4
## Chrysler Imperial 14.7   8  0    3    4
## Dodge Challenger 15.5   8  0    3    2
## AMC Javelin   15.2   8  0    3    2
## Camaro Z28    13.3   8  0    3    4
## Pontiac Firebird 19.2   8  0    3    2
## Hornet 4 Drive 21.4   6  0    3    1
## Hornet Sportabout 18.7   8  0    3    2
## Valiant       18.1   6  0    3    1
## Merc 240D     24.4   4  0    4    2
## Merc 230      22.8   4  0    4    2
## Mazda RX4     21.0   6  1    4    4
## Mazda RX4 Wag 21.0   6  1    4    4
## Merc 240C     24.4   4  0    4    2
## Duster 360    14.3   8  0    3    4

## [1] 19.2 19.2

## [1] 19.2

## [1] 20.09062

## [1] 19.2
```

On note que la médiane est beaucoup plus petite que la moyenne, ce qui suggère une asymétrie dans la distribution de la variable `mpg`.

Regardons l'histogramme de `mpg` pour y voir plus clair:

```
hist(mtcars$mpg)
```



Il s'agit d'une distribution qui lui pourrait qualifier de bimodale, puisqu'on ne voit pas la forme d'une seule cloche mais plutôt deux modes (ou plus de données sont concentrées (c.-à-d. autour de 50-60 ainsi qu'autour de 120-140).

On voit aussi que 5 semble effectivement y avoir un mode (le plus grande fréquence autour de 5).

Voyons ce qui se passe lorsque la variable que l'on veut calculer la moyenne contient des données manquantes:

```
mean(mtcars$cyl[c(1:10,NA)])

## [1] NA
```

Notez qu'il existe une façon de calculer la moyenne en laissant tomber automatiquement toutes les valeurs manquantes, grâce à l'option `na.rm=TRUE`:

```
mean(mtcars$cyl[c(1:10,NA)], na.rm=TRUE)

## [1] 6.2
```

L'option `na.rm=TRUE` peut aussi être utilisée à l'intérieur d'autres fonctions communes en R, comme lors du calcul de la variance pour une variable avec des données manquantes (voir plus bas).

Variance et écart-type

La variance et l'écart-type ne sont pas des mesures de centre, mais plutôt des mesures de dispersion d'une variable.

Soit une variable aléatoire X et n réalisations de cette variable (données par $x_1, ..., x_n$), et soit $\bar{x}$ la moyenne échantillonnale, on a

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

où s^2 dénote la variance échantillonnale. Ainsi, on a aussi

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

qui représente l'écart-type échantillonnal.

Observons par example la distribution théorique d'une variable aléatoire de la Normale(moyenne=1, variance=2) et une variable aléatoire de la Normale(moyenne=1, variance=20):

```
par(mfrow=c(1,2))
plot(dnorm(seq(-20,22,0.01),1,sqrt(2)), ylab="dnorm(x,1,sqrt(2))",xlab="x",type="l")
plot(dnorm(seq(-20,22,0.01),1,sqrt(20)), ylab="dnorm(x,1,sqrt(20))",xlab="x",type="l")
```

On remarque que la distribution Normale avec variance égale à 20 a des observations plus dispersées alors que la majorité des données sont concentrées près du centre pour la distribution de variance 2. Notez que pour visualiser cette comparaison de façon juste, on doit utiliser la même axe des x et la même axe des y.

Dans le jeu de données `Sepal.Length`, examinons la variance échantillonnale de quelques-unes des variables au format numérique:

```
mtcars[order(-mpg),c(1:2,9:11,16:17)]

## [1] 15.19866

## [1] 3.9

## [1] 15.36775

## [1] 0.4375000

## [1] 0.4166667

## [1] 1.615185

## [1] 2.374684

## [1] 0.5625000

## [1] 261.7017

## Warning: This is a deprecated quick access shortcut.
## Use:mtcars[order(-mpg), c(1:2,9:11,16:17)]
## instead.
```

On note que certaines variables, comme `wt`(`cyl`) et `qsec`(`am`) présentent des très grandes variances. Regardons le résumé de la variable `wt`:

```
summary(wt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 1.513   2.581   3.325   3.217   3.610   5.424
```

Le résumé contient la valeur minimum prise par la variable dans l'échantillon (`1.5`), la valeur maximum prise par la variable dans l'échantillon (`5.5`) ainsi, la médiane et la moyenne (respectivement 7.68 et 910.8), quelle différence) ainsi que les quartiles inférieur, Q1 et Q3.

Qu'est-ce que sont ces quartiles Q1 et Q3? Eux, en passant, prennent les valeurs -1 et 1460 le dans notre échantillon. Q1 bien le représentent les valeurs **au-dessous** Jusqu'à 25% et 75% des valeurs de `qsec` se retrouvent, dans l'échantillon, respectivement.

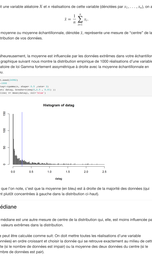
Une boîte à moustaches peut nous aider à visualiser les quartiles Q1, Q3 et la médiane:

```
boxplot(wt,horizontal=TRUE)
```

Ici, il est un peu difficile de distinguer le quartile Q1 qui est égal à -1 et représenterait la limite inférieure de la boîte grise pâle. Le quartile Q3, quant à lui, représente la limite supérieure de la boîte grise. La médiane est aussi confirmée dans la ligne du graphique.

Regardons maintenant la boîte à moustaches de `mpg`:

```
boxplot(mpg,horizontal=TRUE)
```




```
summary(mpg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.4    15.3    19.2    20.1    22.8    33.9
```

Pour se rassurer quant au diagramme, on peut ajouter sur le diagramme le minimum, le maximum, Q1, Q3 et la médiane comme suit:

```
boxplot(mpg,horizontal=TRUE)
abline(v=10.4, col="red")
abline(v=33.9, col="red")
abline(v=15.3, col="blue")
abline(v=22.8, col="blue")
abline(v=19.2, col="green")
```


Étendue interquartile

L'étendue interquartile est égale à $Q_3 - Q_1$ et représente donc la différence entre les deux quartiles Q1 et Q3. Elle nous donne une idée de l'étendue sur laquelle on retrouve 50% des données "centrales".

En R, il existe une fonction pour calculer l'étendue interquartile (EQR, IQR en anglais):

```
IQR(mtcars$mpg)

## [1] 7.5
```

Exercice

L'exercice suivant utilise le jeu de données `swiss` disponible en R. Vous pouvez le télécharger dans R en utilisant la commande suivante:

```
data(swiss)
```

1. Rapportez la moyenne, la médiane et la variance de chacune des deux variables `Sepal` et `Height` du jeu de données.
2. Produisez un diagramme à moustache des deux variables séparément grâce à la fonction

`boxplot` en R. Discutez des résultats. Est-ce que la distribution semble symétrique et unimodale? Est-ce que la boîte grise semble bien répartie de chaque côté de la médiane?

3. En utilisant les méthodes vues dans la capsule 3, calculez l'indice de masse corporelle des patients dans ce jeu de données. Ensuite, produisez le diagramme à moustache de l'IMC et discutez de l'utilisation de la moyenne ou de la médiane comme mesure de centre dans ce contexte. Laquelle de la moyenne ou de la médiane représenterait la meilleure mesure de centre à utiliser pour décrire la distribution, selon vous? Pourquoi?