R Notebook

Correction TD 1

Exercice 1

1.

La population étudiée ici est un ensemble de pays d'Europe, dans lequel chaque individu est un pays. La variable étudiée correspond à la température moyenne observée en 2018.

2.

Pour calculer la moyenne de l'échantillon, il suffit de calculer (ici $n = 14$) :

$$\bar{x} = \frac{1}{14}\Big(15.2 + 10.5 + \cdots + 9.3\Big) \approx 10.543$$

3.

On calcule tout d'abord

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{14}\Big(15.2^2 + 10.5^2 + \cdots + 9.3^2\Big) \approx 126.72$$

On en déduit

$$\bar{s}_b^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \approx 126.72 - 10.543^2 \approx 15.57$$

4.

Pour calculer la médiane, on pose $p = \frac{1}{2}$, et on applique la formule :

$$j = \Big\lfloor (n-1)p \Big\rfloor = \Big\lfloor (14-1)\frac{1}{2} \Big\rfloor = \Big\lfloor \frac{13}{2} \Big\rfloor = \lfloor 6.5 \rfloor = 6$$

et $r = (n-1)p - j = 6.5 - 6 = 0.5$.

Il reste à classer les données de la plus petite à la plus grande pour trouver la 6-ième :

1	2	3	4	5	6	7	8	9	10	11	12	13	14
4.3	5.7	5.9	6.6	9.1	9.3	10.3	10.5	10.7	12.3	15	15.2	15.2	17.5

Donc $x_{j+1} = x_7 = 10.3$, et $x_{j+2} = x_8 = 10.5$, et enfin

$$\bar{q}_{1/2} = x_{j+1} + r * (x_{j+2} - x_{j+1}) = 10.3 + 0.5 * (10.5 - 10.3) = 10.3 + 0.5 * 0.2 = 10.4$$

La médiane vaut donc 10.4, et il y a bien la moitié des données en dessous, et la moitié au dessus.

On peut faire le même calcul pour trouver $\bar{q}_{\frac{1}{4}} = 7.225$ et $\bar{q}_{\frac{3}{4}} = 14.325$.

5.

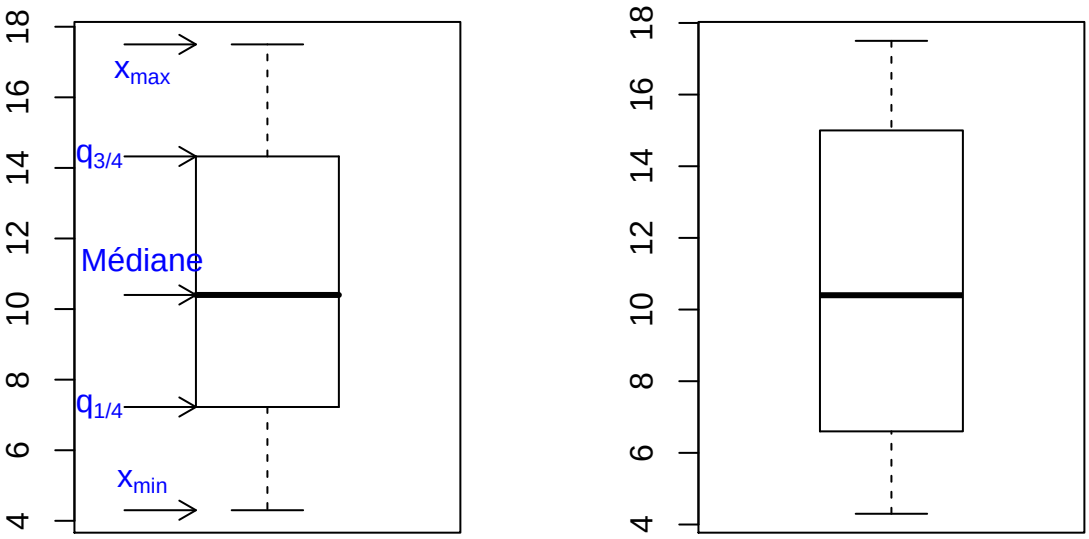
On doit d'abord calculer $e = \bar{q}_{\frac{3}{4}} - \bar{q}_{\frac{1}{4}} = 14.325 - 7.225 = 7.1$, puis

$$x_{min} = \max(\min_i(x_i), \bar{q}_{\frac{1}{4}} - 1.5e) = \max(4.3, 7.225 - 1.5 * 7.1) = 4.3,$$

et

$$x_{max} = \min(\max_i(x_i), \bar{q}_{\frac{3}{4}} + 1.5e) = \min(17.5, 14.325 + 1.5 * 7.1) = 17.5$$

On peut dès lors représeter le boxplot de x :



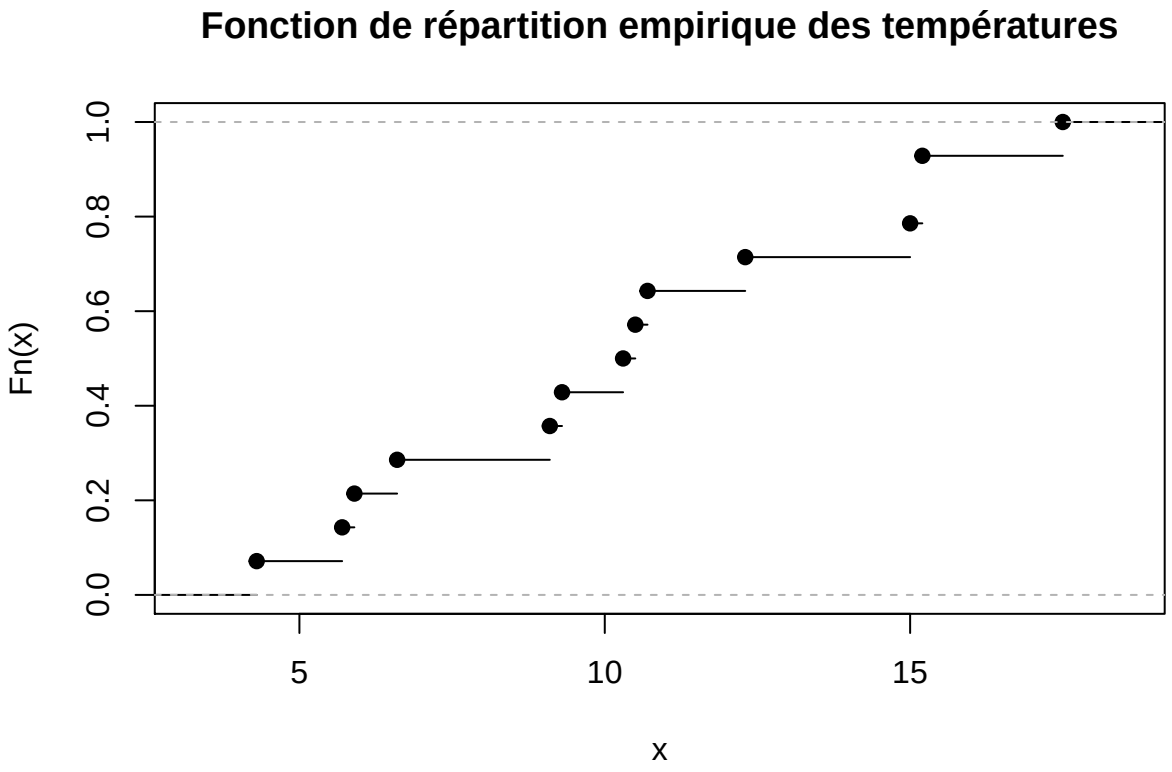
6.

Il suffit de compter combien de températures dans le tableau sont inférieures ou égale à 13, ici 10. On a donc

$$\bar{F}(13) = \frac{10}{14} \approx 0.714.$$

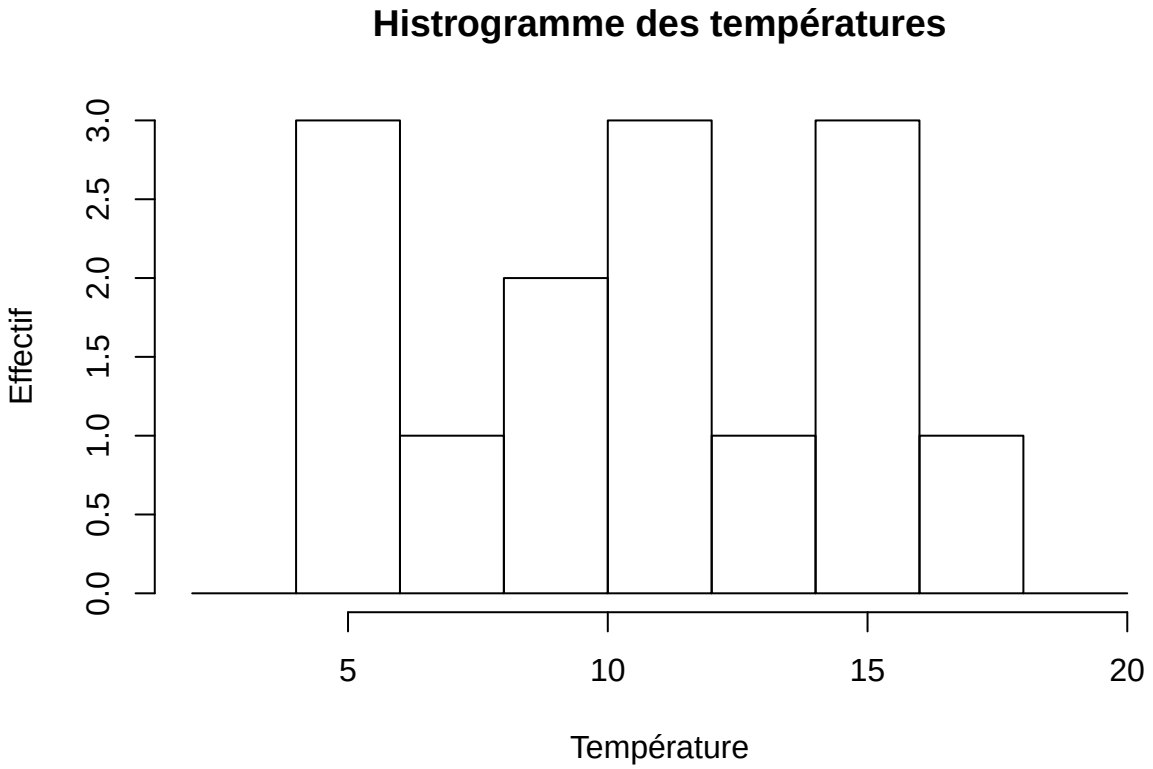
7.

Pour représenter la fonction de répartition empirique, il faut remarquer qu'elle augmente de $1/n$ à chaque point de l'échantillon. On obtient alors



8.

Il suffit de regroupper les données par classe, et de compter les effectifs de chaque classe. On obtient alors



Exercice 2.

1.

La variable aléatoire correspondant à un dé équililbré est une variable aléatoire discrète à valeurs dans $\{1,2,3,4,5,6\}$ dont la loi est donnée par

$$\forall k \in \{1,2,3,4,5,6\}, \mathbb{P}(X=k) = \frac{1}{6}.$$

On peut construire aisément un espace de probabilité sur lequel on va construire X en prennant :

- $\Omega = \{1,2,3,4,5,6\}$
- $\mathcal{A} = \mathcal{P}(\Omega)$
- $\forall A \in \mathcal{A}, \mathbb{P}(A) = \frac{Card(A)}{6}$, où $Card(A)$ désigne le nombre d'éléments dans A.
- $\forall \omega \in \Omega, X(\omega) = \omega$.

2.

On calcule

$$\mathbb{E}[X] = \frac{1}{6} * 1 + \frac{1}{6} * 2 + \cdots + \frac{1}{6} * 6 = 3.5$$

- puis

$$\mathbb{E}[X^2] = \frac{1}{6} * 1^2 + \frac{1}{6} * 2^2 + \cdots + \frac{1}{6} * 6^2 = \frac{91}{6},$$

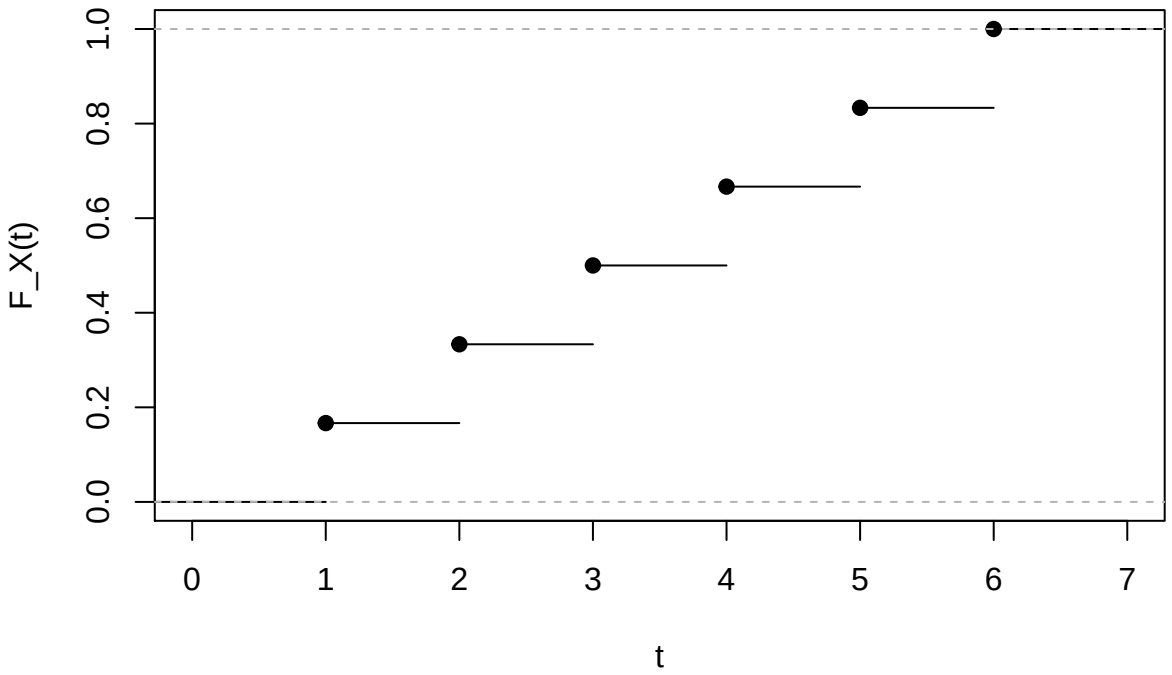
d'où

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{35}{12} \approx 2.92$$

La fonction de répartition de X est donnée par

$$F_X(t) = \begin{cases} 0 \text{ si } t < 1 \\ \frac{1}{6} \text{ si } 1 \leq t < 2 \\ \frac{2}{6} \text{ si } 2 \leq t < 3 \\ \vdots \\ 1 \text{ si } t \geq 6 \end{cases}$$

On peut la représenter :

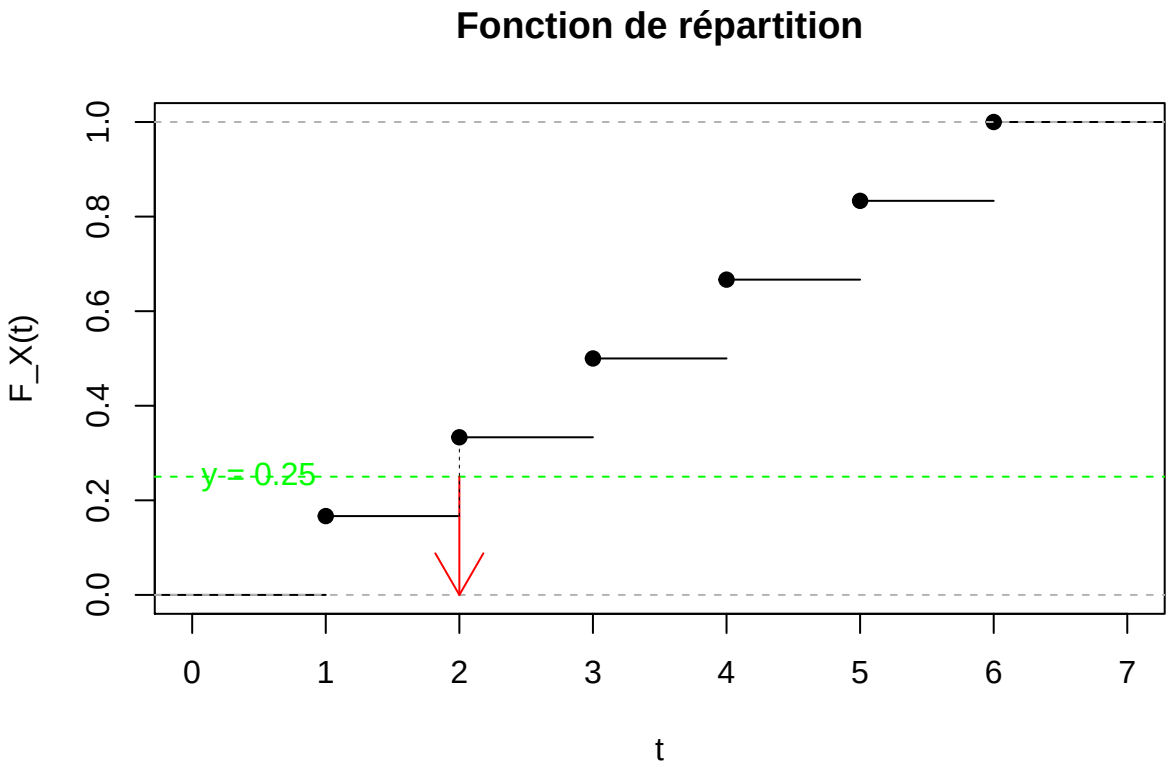


Le quantile théorique d'ordre $1/4$ est défini par $q_{1/4} = \inf\{x \in \mathbb{R}, F_X(x) \geq \frac{1}{4}\}$

Pour le calculer, il suffit d'identifier l'ensemble $\{x \in \mathbb{R}, F_X(x) \geq \frac{1}{4}\}$ qui n'est autre que l'ensemble des points pour lesquels la fonction de répartition est au dessus de 0.25, i.e. ici l'intervalle $[2, +\infty[$. $q_{1/4}$ est le plus petit élément de cet ensemble, donc

$$q_{1/4} = 2.$$

On peut le lire sur le graphique :



3.

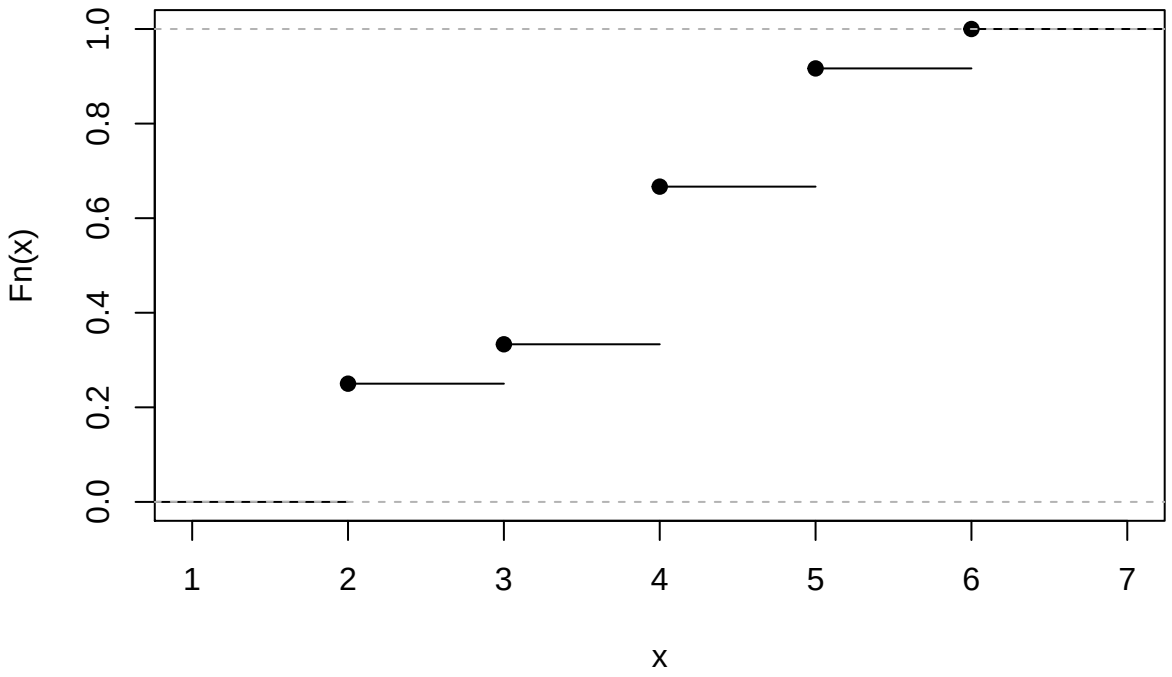
Pour donner la fréquence de chaque résultat, il suffit de diviser les effectifs de chaque résultats par l'effectif total :

Valeur	1	2	3	4	5	6
Fréquence	0	$\frac{3}{12}$	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{3}{12}$	$\frac{1}{12}$

En reprennant les idées de l'exercice 1, on trouve

$\bar{x} \approx 3.8333$ $\bar{s}_b^2 \approx 1.64$ $\bar{q}_{\frac{1}{4}} = 2.75$

Et enfin, la fonction de répartition empirique est donnée par le graphique suivant :



L'objectif de cet exercice est de bien comprendre la différence entre :

- Variable aléatoire et échantillon de réalisations indépendantes de tirages de cette variable
- "Moyenne" (espérance) d'une variable aléatoire et moyenne empirique d'un échantillon
- Variance d'une variable aléatoire, et variance empirique d'un échantillon
- Fonction de répartition théorique d'une variable aléatoire, et fonction de répartition empirique d'un échantillon
- Quantile théorique d'une variable aléatoire, et quantile empirique d'un échantillon...

4.

Cette question est vraiment très importante !! L'objectif est de comprendre qu'à chaque expérience de 12 lancers de dé indépendants, $\bar{x}_{12}$ peut donner des valeurs différentes. Si on répète l'expérience de lancers de 12 dès un grand nombre de fois, certaines valeurs pour $\bar{x}_{12}$ sortirons plus que d'autres.

On peut donc parler de la variable aléatoire $\bar{X}_{12}$ **AVANT** de faire l'expérience, comme le résultat (aléatoire donc) d'une expérience où on lance 12 dès, et on fait la moyenne.

Par définition, $\bar{X}_{12}$ étant de la forme $\frac{S}{12}$ où S correspond à la somme de 12 lancers (donc $S \in [\![12, 72]\!]$), la variable aléatoire $\bar{X}_{12}$ est une variable discrète qui ne peut prendre qu'un nombre fini de valeurs de la forme

$$\frac{12}{12}, \frac{13}{12}, \frac{14}{12}, \cdots, \frac{72}{12}.$$

Pour la construire, on peut par exemple prendre l'espace de probabilité correspondant au lancers de 12 dés équilibrés (à 6 faces) indépendants :

- $\Omega = \{1, 2, \cdots 6\}^{12}$
- $\mathcal{A} = \mathcal{P}(\Omega)$

- $\forall A \in \mathcal{A}, \mathbb{P}(A) = \frac{Card(A)}{6^{12}}$
- $\forall \omega \in \Omega, X(\omega) = \frac{\omega_1 + \cdots + \omega_{12}}{12}$.

On peut calculer certaines probabilités, par exemple, pour $\mathbb{P}(\bar{X}_{12} = 1)$, il suffit de remarquer que la seule façon d'obtenir 1 comme moyenne sur le lancer de 12 dès, c'est que chaque dé donne 1. Donc en notant $D_1, \cdots, D_{12}$ les variables aléatoires correspondants aux 12 lancers de dé, on obtient

$$\mathbb{P}(\bar{X}_{12} = 1) = \mathbb{P}(D_1 = D_2 = \cdots = D_{12} = 1) = \mathbb{P}(D_1 = 1)\mathbb{P}(D_2 = 1) \cdots \mathbb{P}(D_{12} = 1),$$

en utilisant l'indépendance pour passer au produit. On obtient alors

$$\mathbb{P}(\bar{X}_{12} = 1) = \frac{1}{6^{12}} \approx 4.6 * 10^{-10}.$$

De même, pour $\mathbb{P}(\bar{X}_{12} = \frac{13}{12})$, il faut avoir obtenu 11 lancers qui ont donné 1 et un 2, ce qui peut être fait de 12 façons possibles. On trouve

$$\mathbb{P}(\bar{X}_{12} = \frac{13}{12}) = \frac{12}{6^{12}} \approx 5.5 * 10^{-9}.$$

Pour $1.25 = 15/12$, il faut être un peu plus motivé pour compter les possibilités :

- (11*1,4) : 12 façons
- (10*1,2,3) : 132 façons : $2\binom{12}{2}$
- (9*1,2,2,2) : 220 façons : $\binom{12}{3}$

Chaque possibilité étant équiprobable, on obtient

$$\mathbb{P}(\bar{X}_{12} = 1.25) = \frac{12 + 132 + 220}{6^{12}} \approx 1.67 * 10^{-7}.$$

5.

On ne peut pas ici appliquer la formule de l'espérance $\sum_k a_k \mathbb{P}(X = a_k)$, vu qu'on ne connaît pas $\mathbb{P}(X = a_k)$.

Par contre, il suffit d'utiliser la linéarité de l'espérance :

$$\mathbb{E}[\bar{X}_{12}] = \mathbb{E}\left[\frac{1}{12}\sum_{i=1}^{12} D_i\right] = \frac{1}{12}\sum_{i=1}^{12} \mathbb{E}\big[D_i\big] = \frac{1}{12}\sum_{i=1}^{12} 3.5 = 3.5.$$

Exercice 3 :

- La race d'un chien : Qualitative discrète
- Le salaire d'un individu : Quantitative continue
- La marque d'un téléphone portable : Qualitative discrète
- L'âge d'un individu : Quantitative discrète si compté en valeur entière, continue sinon
- Le nombre de députés d'un pays : Quantitative discrète
- La température d'un réacteur nucléaire : Quantitative continue
- Le nombre de frères et soeurs d'un individu : Quantitative discrète

Exercice 4 :

On numérote les graphiques de gauche à droite et de haut en bas.

1)

L'histogramme de revenus nets indique, par tranches de revenus, le nombre de foyers ayant un revenu dans chaque tranche. Il est pertinent pour une variable continue, et donne une idée de la répartition des revenus par foyer. Par exemple, on remarque que la majorité des foyers est au dessous de 3000 euros par adulte, et que très peu de personnes gagnent plus de 4000.

2)

Le barplot de revenus indique, pour chacun des 50 foyers, le revenu correspondant. Ce graphique est difficile à lire et de peu d'utilité. Si l'on souhaite représenter cela, il est beaucoup plus intéressant de classer les foyers par revenus, avant de procéder à un barplot, ce qui permet de visualiser la fonction quantile facilement.

3)

Ce diagramme circulaire représente, pour chaque foyer, une aire correspondant à la proportion de revenu gagné parmi le revenu total gagné par les 50 foyers. Ici, ce graphique semble très peu pertinent, vu que le total n'a pas vraiment de sens. Le seul cadre dans lequel ce graphique aurait du sens serait par exemple pour illustrer qu'à l'échelle mondiale, une part non négligeable des richesses est détenue par peu d'individus (auquel cas on s'interesse bien à une proportion de revenu/richesses parmis un total qui a du sens.)

4)

Le boxplot des revenus est une représentation synthétique de la position et de la disparité entre revenu (dispersion des variables). Ici, on peut lire très directement :

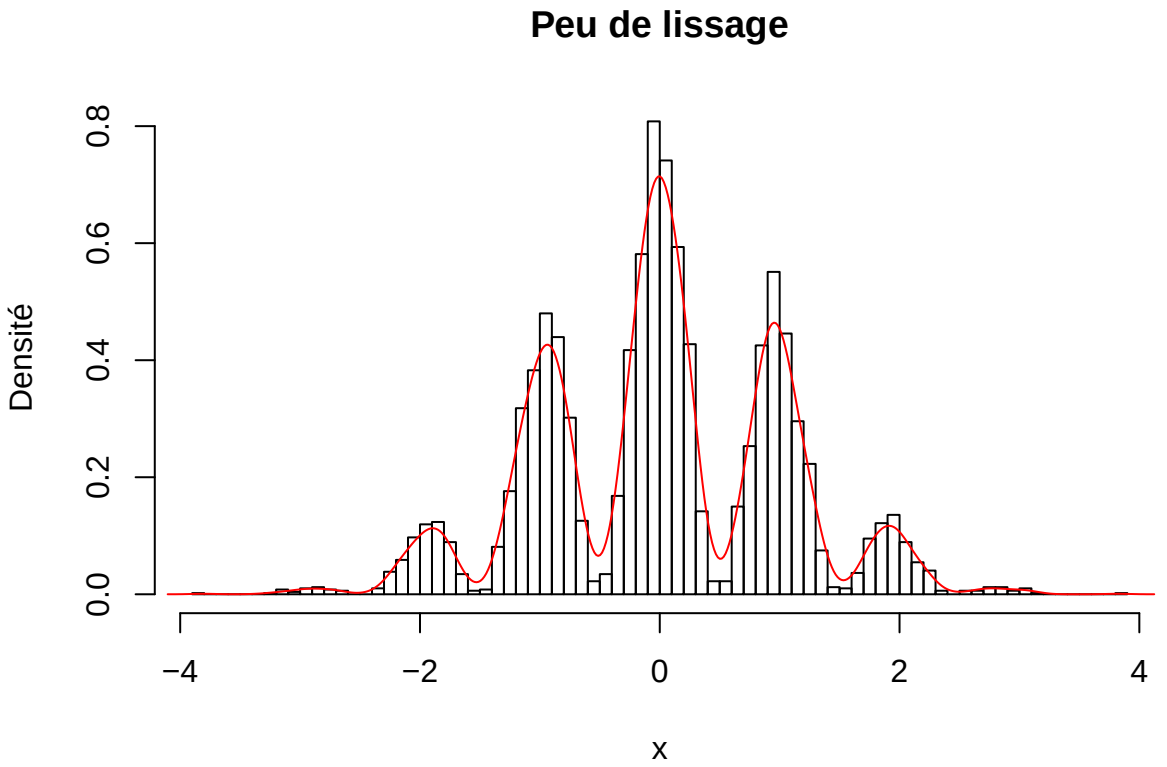
- que la médiane est proche de 1600 euros sur nos 50 foyers (pour info, la médiane du niveau de vie en 2017 en Auverge-Rhône-Alpe était donnée à 1820 euros par mois par **l'INSEE**)
- que les "bas" salaires sont assez "tassés"
- que les "hauts" salaires sont très dispersés, i.e. qu'il y a dans l'échantillon des salaires extrèmes, élevés d'un point de vue statistique au dessus de $q_{3/4} + 1.5(q_{3/4} - q_{1/4})$. Ce graphique est donc pertinent

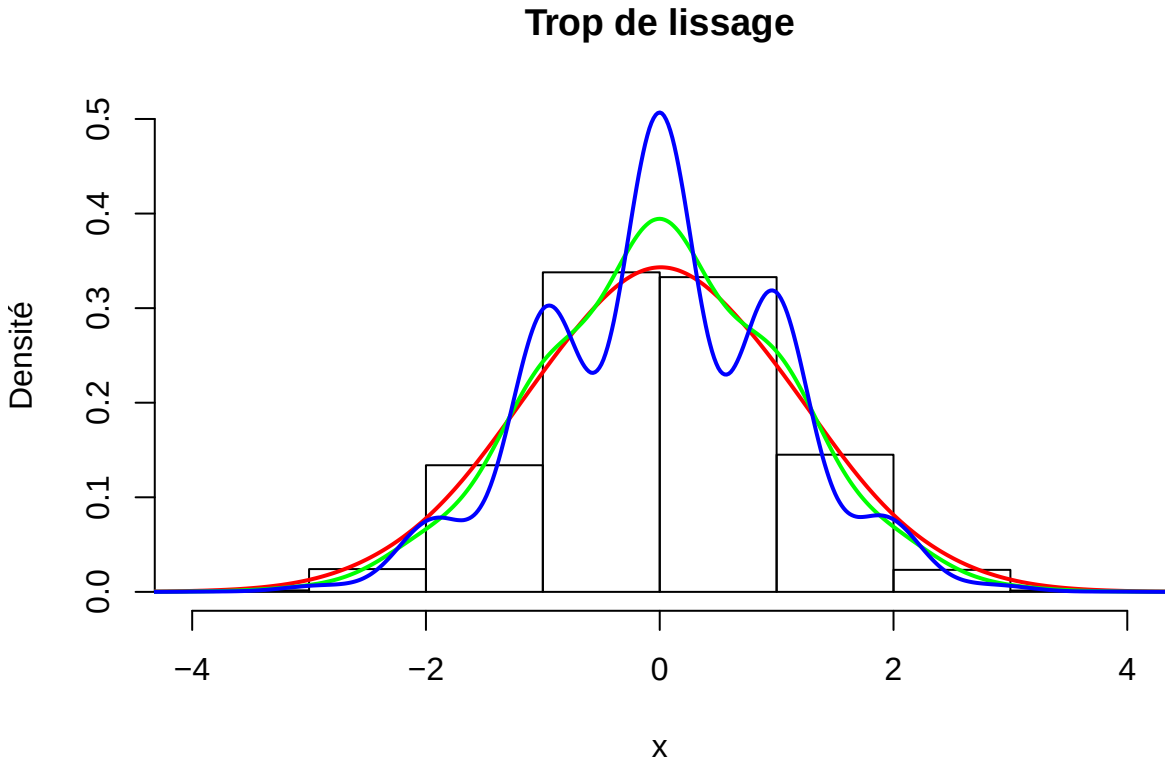
5)

La fonction de répartition indique, pour chaque revenu, quelle est la proportion qui gagne moins. C'est souvent une représentation pertinente, en particulier pour une variable continue, car on peut observer aisément les revenus qui concentrent beaucoup de foyers "la courbe monte vite", et ceux qui en concentrent beaucoup moins "la courbe monte lentement". On peut faire ici les mêmes commentaires que pour l'histogramme.

6)

Ce plot a pour objectif de vous montrer que parfois on fera des représentations peu usuelles. Ici, plutôt que de représenter la densité, on la représente verticale, et symétrisée, afin d'obtenir cette forme de "violon". Comme toute représentation de la densité, ce graphique est construit en "lissant l'histogramme", et donne sensiblement les mêmes indications, même s'il faut être prudent car le "lissage" peut "cacher" des choses importantes (tout comme un histogramme avec des barres trop larges), par exemple





7)

Le dernier plot représente juste une variable en fonction de l'autre. Ce plot est très utile, car il permet de représenter la relation entre les deux. En particulier, on semble voir du premier coup d'oeil qu'il y a peu de

revenus très faibles avec un nombre d'années d'études post 16 ans très élevé, et à l'inverse, peu de revenus élevés avec un nombre d'années d'études faible. Cela n'indique absolument rien sur le sens de la causalité, mais on peut imaginer que le coût élevé des études est un frein immense pour les faibles revenus, et qu'à l'inverse, un revenu élevé permet par exemple de financer des écoles privées sans contraintes d'admission.

Au passage, sans faire de calculs, on a l'impression que les données sont "corrélées", i.e. un peu "alignées" sur une droite ni verticale ni horizontale.

8)

Comme pour le précédent diagramme circulaire, cela ne fait aucun sens de regarder le nombre d'années d'étude total de l'échantillon, et une proportion parmi celui ci. (on ne peut pas penser "il a pris toutes les années d'études..."). Ce plot n'est donc pas du tout pertinent.

9)

Ce barplot représente un barplot d'effectifs et se lit de la façon suivante : parmi les 50 foyers, 6 ont des enfants ayant un niveau d'étude (post 16 ans) de 1 an, 12 de 2 ans...

Ce graphique est utile car il résume en un coup d'oeil les données (donc pertinent car la variable est discrète).

10)

De même, le diagramme circulaire des effectifs est pertinent, puisque le total représente les 50 foyers, et que cette fois l'angle représente la proportion d'enfants ayant fait 1 an, 2 ans... d'études.

11)

Ce boxplot indique la dispersion, mais c'est un peu délicat pour une variable ayant aussi peu de modalités (ici 9 différentes), donc on préfèrera les deux graphiques précédents.

12)

La fonction de répartition est moins rapide à lire pour une variable discrète avec peu de modalités, mais est toujours pertinente. Pour cet exemple, je trouve que le barplot ou le pieplot (camembert) donnent plus d'informations visuelles.

13)

On évitera de réprésenter des densités pour des variables discrètes. Ce plot n'est donc pas pertinent, et à éviter tant que possible (exception faite des cas où il y a beaucoup de modalités différentes et ordonnées, mais dans ce cas là, on pourrait modéliser la variable comme continue...)

Exercice 5 :

1)

Si on regarde globalement, le traitement B est meilleur. Mais si on distingue par sévérité du calcul, alors le A est meilleur pour les petits calculs **et** pour les gros. C'est la que réside le paradoxe, si A est meilleur pour chaque sévérité, on s'attend à ce qu'il soit meilleur globalement.

2)

En regardant les chiffres en détails, on peut se rendre compte que le traitement A a été majoritairement prescrit pour les gros calculs alors que le B est majoritairement prescrit pour des petits. On remarque aussi que les petits calculs ont un taux de guérison plus élevé que les gros (pour chaque traitement). Le traitement A étant utilisé majoritairement pour des cas plus sévères, même s'il est meilleur que le B, il aura un taux global de guérison plus faible.

On peut imaginer que les docteur.e.s avaient plus confiance dans le traitement A, et que c'est la raison pour laquelle ils/elles l'ont prescrit pour les cas les plus difficiles.

3)

Ce paradoxe motive l'utilisation de **randomisation**, i.e. dans ce cas là, on choisirait au hasard le traitement donné à chaque patient, afin d'éviter d'influencer les résultats par le choix du traitement. Comme il existe de nombreux biais (par exemple la croyance du/de la médecin ou du/de la patient · e dans la performance d'un traitement), on procède souvent à une étude "en double aveugle" où ni la/le patient · e ni la/le médecin ne sait quel traitement est utilisé (ou placébo vs traitement). Plus d'infos **ici** par exemple.

Enfin, cela n'est parfois pas possible de procéder à une randomisation (par exemple si on veut regarder l'effet de la cigarette sur la santé, on ne peut pas tirer des gens au hasard et leur demander de se mettre à fumer), dans ce cas là, il existe des méthodes (compliquées !) pour prendre en compte les "facteurs de confusion" (par exemple que les fumeurs ont aussi plus tendance à consommer d'autres drogues que les non fumeurs, et on veut différencier les effets de la cigarette et des autres drogues éventuellement consommées).

Exercice 6 :

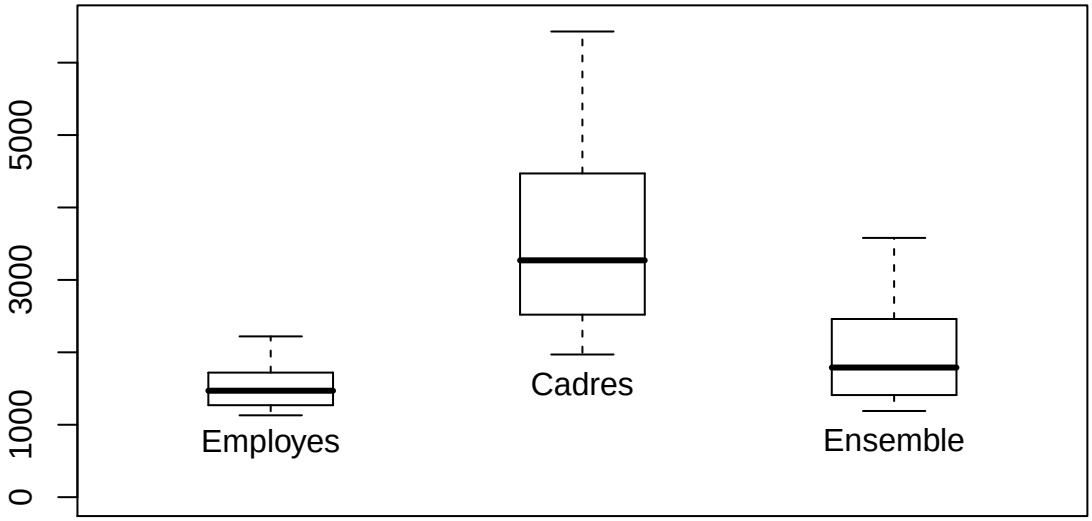
Rien ne va !!!

1) Il y a déjà un biais induit par "on veut montrer". On ne procède pas à l'expérience pour répondre à la question "est-ce que", mais avec un objectif précis. Cela peut induire des biais de confirmation, et des conflits d'intéret (si jamais on y gagne de l'argent par exemple)
2) Il y a deux problèmes à "interoger des personnes devant la gare Partdieu" : l'échantillon n'est vraisemblablement pas représentatif (en particulier, il doit y avoir une probabilité plus grande que ces personnes n'habitent pas à Lyon qu'une personne tirée au hasard dans Lyon intramuros). Le second problème est qu'il n'est pas précisé que les personnes sont choisies "au hasard"
3) On ne peut pas recommencer l'expérience quand les résultats ne sont pas satisfaisant ! Cela fausse évidemment les statistiques
4) Cette fois, il est précisé que les personnes sont intérogées "au feeling", donc pas au hasard
5) On ne peut pas non plus stopper un expérience quand le résultat nous convient, il faut prévoir à l'avance le nombre d'individus que l'on va intégrer dans l'étude
6) Un échantillon de 9 semble bien petit !
7) Même s'il n'y avait aucun des problèmes précédents, on ne peut pas conclure seulement en calculant le ratio, il est nécessaire d'effectuer un raisonnement mathématique (et c'est l'objet de cette UE) pour se prémunir de la variabilité dans les données (avoir un résultat "par chance" sur cette expérience)

De façon générale, retenez que **le protocole et le plan d'analyse statistique doivent être écrits avant l'experience**

Exercice 7 :

Catégorie	Ensemble	Cadres	Employés
1eme décile	1190	1970	1130
1er quartile	1410	2520	1270
Médiane	1790	3270	1470
3eme quartile	2460	4470	1780
9eme décile	3580	6430	2220



Exercice 8 à 14 :

Exercices bonus pour la culture ou pour s'entraîner, non corrigés ici.

A voir avec le/la binôme si cela été vu en TD, où poser une question en début/fin de cours/TD (pour une question précise, vous pouvez aussi m'envoyer un mail.).