Corrigé 1

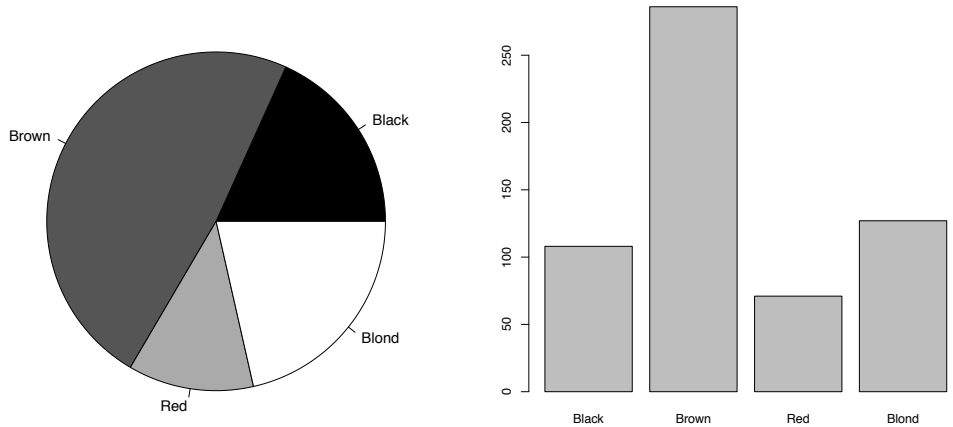
Exercice 1. (a) Il y a 592 étudiants.

(b) Dans cette étude, les variables sont qualitatives nominales. Cependant, on pourrait faire une étude plus avancée qui étudierait les couleurs de cheveux et d'yeux comme un continuum. On pourrait donc faire une autre étude où ces mêmes variables sont continues.

(c) La plupart des étudiants ayant participé à l'enquête ont les cheveux bruns. La couleur la moins fréquente est roux. Les fréquences absolues et relatives sont :

Couleur	**Fréquence absolue**	**Fréquence relative**
Noir	108	0.182 (18.2 %)
Brun	286	0.483 (48.3 %)
Roux	71	0.120 (12.0 %)
Blond	127	0.215 (21.5 %)

On peut représenter les résultats à l'aide des graphiques suivants :



Pour une fois, le choix du "pie-chart" peut se défendre, car il permet de voir clairement que la moitié de la population est brune.

(d) Examinons les fréquences relatives des couleurs des yeux parmi les étudiants avec les cheveux noirs et parmi les étudiants avec les cheveux blonds. Les fréquences relatives des couleurs des yeux parmi les étudiants avec les cheveux noirs sont :

Brun	Bleu	"Hazel"	Vert
63.0 %	18.5 %	13.9 %	4.6 %

Par contre, les fréquences relatives des couleurs des yeux parmi les étudiants avec les cheveux blonds sont :

Brun	Bleu	"Hazel"	Vert
5.5 %	74.0 %	7.9 %	12.6 %

Puisque les fréquences relatives semblent différentes entre les deux groupes (cheveux noir vs cheveux blonds), il semble que la couleur des yeux et la couleur des cheveux des étudiants ayant participé à l'enquête sont liées. Dans quelques semaines nous allons voir comment faire un test formel pour cela.

Exercice 2. (a) Quantitative.

(b) $\bar{x} = 165.067$ cm,
$s_x = 11.423$ cm.

(c) $147, 150, 152, 155, 157, 160, 163, 165, 168, 170, 173, 175, 178, 180, 183$.

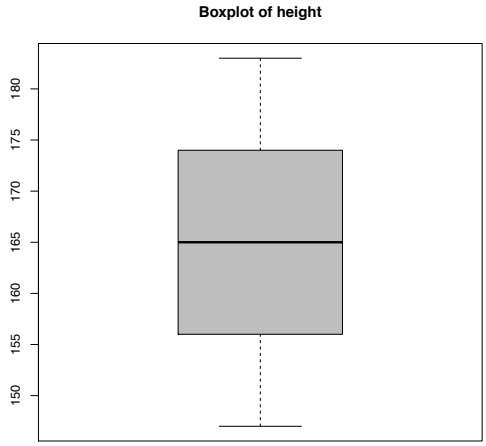
(d) $x_{(1)} = 147$ cm,
$x_{(15)} = 183$ cm,
$med(x) = x_{(8)} = 165$ cm,
$\hat{q}_x(25\,\%) = x_{(4)} = 155$ cm,
$\hat{q}_x(75\,\%) = x_{(12)} = 175$ cm,
$\hat{q}_x(30\,\%) = x_{(5)} = 157$ cm.
L'idée des quantiles est qu'ils partagent les données en deux parties d'une manière spéciale. Par exemple, le quantile d'ordre 30 %, $\hat{q}(30\,\%)$, est une valeur telle qu'environ 30 % des données sont inférieures à cette valeur et environ 70 % des données sont supérieures à cette valeur.

(e) $IQR(x) = \hat{q}_x(75\,\%) - \hat{q}_x(25\,\%) = 175 - 155 = 20$ cm.

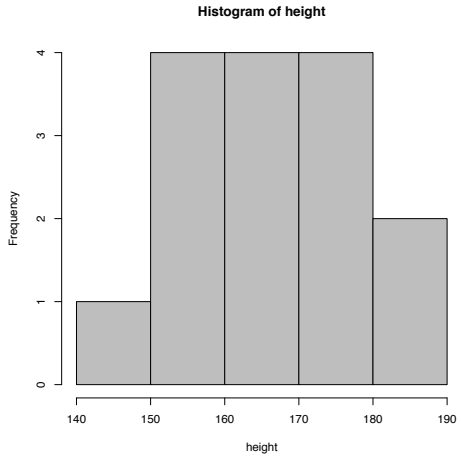
(f) Un boxplot montre d'une certaine manière "où sont les données". La région grise contient 50 % des données centrales. La région entre les limites de la "moustache" montre où les données typiques sont attendues. Les données en dehors de cette région (s'il y en a) pourraient être aberrantes. Les extrémités des moustaches du boxplot se calculent de la façon suivante :

$$\min\{x_i : x_i \geq \hat{q}_x(25\,\%) - C\} = \min\{x_i : x_i \geq 155 - 30\} = x_{(1)} = 147$$

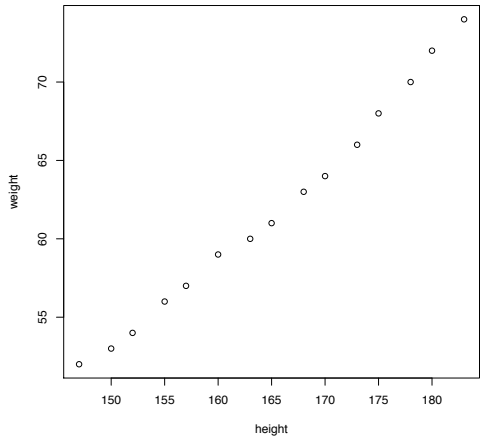
$$\max\{x_i : x_i \leq \hat{q}_x(75\,\%) + C\} = \max\{x_i : x_i \leq 175 + 30\} = x_{(15)} = 183.$$



(g) Un histogramme donne, dans un sens, plus d'informations que le boxplot, même s'il contient moins d'informations que les données initiales. Dans ce cas, il suggère que la distribution est plutôt symétrique autour de la moyenne 165 et que toute sauf trois personnes ont la taille entre 150 et 180cm.



(h) Le poids des femmes ayant participé à l'enquête augmente avec la taille. Les deux caractéristiques semblent donc liées. Dans quelques semaines nous allons voir comment faire un test formel pour cela.



Exercice 3. (a) Pour la moyenne du nouvel échantillon, on trouve

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(a\,x_i + b) = b + a \times \frac{1}{n}\sum_{i=1}^{n} x_i = a\bar{x} + b.$$

Pour la médiane, on utilise le fait que $x \mapsto ax + b$ est une transformation linéaire, donc monotone. Ainsi, pour $a > 0$, cette transformation préserve l'ordre, c'est-à-dire $y_{(k)} = a\,x_{(k)} + b$ quel que soit k. Il est alors facile de vérifier que $med(y) = a\,med(x) + b$ lorsque n est pair ou impair. D'autre part, si $a < 0$, la transformation inverse l'ordre : $y_{(1)} = a\,x_{(n)} + b$, $y_{(2)} = a\,x_{(n-1)} + b$ et ainsi de suite. Si n est impair on a $y_{((n+1)/2)} = a\,x_{((n+1)/2)} + b$ et la propriété est vérifiée. Si n est pair, on aura $y_{(n/2)} = a\,x_{(n/2+1)} + b$ et $y_{(n/2+1)} = a\,x_{(n/2)} + b$, et donc $med(y) = a\,\frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) + b$ comme voulu.

Pour l'écart-type, on a

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a\,x_i + b - a\,\bar{x} - b)^2} = \sqrt{a^2\,\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = |a|\,s_x.$$

(b) En utilisant les résultats de l'exercice précédent ainsi que les résultats que l'on vient de démontrer, nous avons :

$$\begin{aligned} \bar{y} &= a\,\bar{x} + b \;=\; 1/2.54 \cdot 165.067 \;=\; 64.987; \\ med(y) &= a\,med(x) + b \;=\; 1/2.54 \cdot 165 \;=\; 64.961; \\ s_y &= |a|\,s_x \;=\; 1/2.54 \cdot 11.423 \;=\; 4.497. \end{aligned}$$

(c) Par lieu, on entend un nombre autour duquel se trouvent les données. Si les données sont décalées vers la gauche ou vers la droite, le lieu est décalé dans la même direction et de la même distance.

(d) Par échelle, on entend l'étendue recouverte par (la majorité) des données. Si les données sont proches l'une de l'autre, la dispersion est petite. Si les données s'éloignent l'une de l'autre, la dispersion augmente. La dispersion est une quantité relative. Par exemple, 4 et 5 kilogrammes sont aussi proche que 4 000 et 5 000 grammes. Pourtant, l'écart-type des données en grammes va être 1 000 fois plus grand que l'écart-type de mêmes données en kilogrammes. Et enfin, si la différence d'un kilogramme est grande ou petite dépend de la situation, par exemple si l'on pèse des bébés ou des adultes.

(e) L'écart inter-quartile et l'étendue sont des statistiques d'échelle tandis que des quantiles sont des statistiques de lieu.