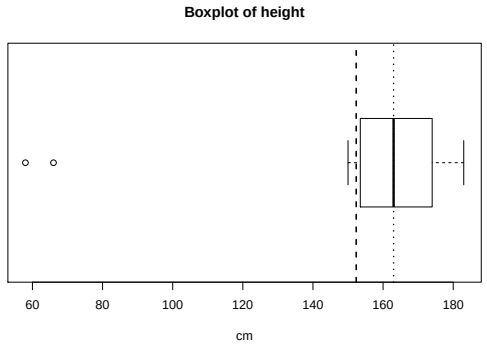
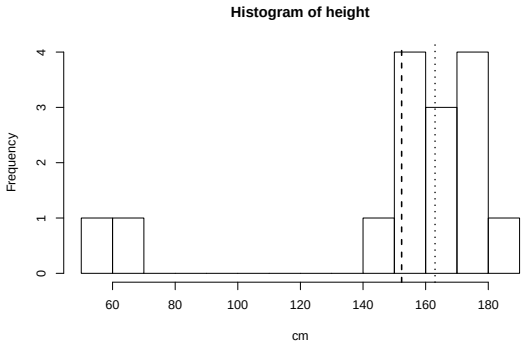
Corrigé 2

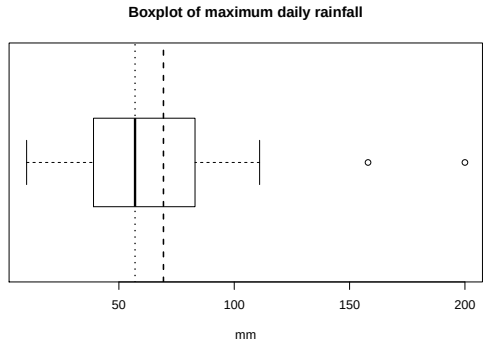
Exercice 1. (a) En traçant la moyenne (152.3 cm) et la médiane (163 cm) sur les graphiques, on voit que la moyenne est décalée vers les deux observations aberrantes. Ici on peut se permettre de dire que ces deux observations sont aberrantes, car il est rare de voir une femme ayant une taille inférieure à 80 cm. En fait, ces données sont les mêmes que celles de l'exercice 2 de la série 1, sauf qu'on a oublié de changer l'unité de mesure des deux premières observations ; celles-ci sont donc présentées en pouces plutôt qu'en centimètres. Il y a donc deux erreurs de mesure dans les données (souvent, les valeurs aberrantes sont dues à des erreurs). La tendance centrale des tailles est donc beaucoup mieux représentée par la médiane.

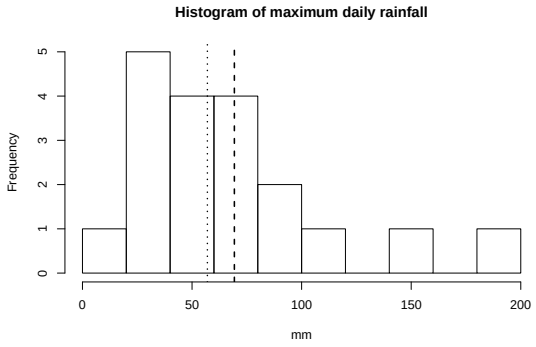




(b) La moyenne est beaucoup plus influencée par un changement d'une petite partie des données que l'est la médiane. Il faut cependant noter que les deux observations aberrantes sont vraiment différentes du reste des données ; un petit changement dans les données n'aurait pas eu un effet si grand sur la médiane.

(c) L'écart inter-quartile. En général, si l'on suspecte la présence de valeurs aberrantes, on utilise la médiane et l'écart inter-quartile, plutôt que la moyenne et l'écart-type pour caractériser la distribution des données.





Exercice 2. (a) En traçant la moyenne (69.32 mm) et la médiane (57 mm) sur les graphiques, on voit que la moyenne est décalée vers les deux observations atypiques, comme dans l'exercice 1. Par contre, dans ce cas-ci, il ne s'agit pas d'une erreur dans les données. Même si la tendance centrale des maxima des pluies journalières typiques est mieux représentée par la médiane, la différence entre la moyenne et la médiane attire notre attention sur le

fait qu'il y eu deux années ayant eu une journée de novembre très pluvieuse. En regardant l'histogramme, on peut aussi constater que les données à droite de la médiane sont plus dispersées que les données à gauche. Cela nous indique que la probabilité d'observer une quantité de pluie journalière beaucoup plus grandes que la médiane n'est pas aussi petite que l'on pourrait le penser (elle est non-négligeable).

(b) Dans l'exercice 1, les données atypiques sont en fait des erreurs dans les données, tandis que dans l'exercice 2, ce sont simplement des observations moins fréquentes. En conséquence on tire des conclusions différentes dans les deux situations. Même s'il n'est pas toujours possible de détecter des erreurs dans les données (comme nous l'avons fait dans l'exercice 1), il est important de bien comprendre les données, et de se demander si les données qui nous paraissent atypiques sont des erreurs ou bien des valeurs correctes moins fréquentes.

Exercice 3. (a) Chaque proportion est entre 0 et 1, et la somme est bien égale à 1.

(b) Le pourcentage des femmes divorcées ou jamais mariées est égal à $0.071 + 0.353 = 0.424$.

(c) Le pourcentage des femmes qui ne sont pas mariées est égal à $1 - 0.574 = 0.426$.

Exercice 4. (a) On a vu dans le cours que pour la distribution $\mathrm{N}(\mu, \sigma^2)$, le pourcentage d'observation dans $[\mu - \sigma, \mu + \sigma]$ est 68 %. Ici, avec $\mu = \bar{x} = 176.6$ et $\sigma = s_x = 7.99$, cet intervalle est $[168.61, 184.59]$ et la fréquence relative vaut 63.3 %, ce qui est relativement proche de la proportion attendue sous la distribution normale.

(b) On standardise par la transformée $x \mapsto (x - \bar{x})/s_x$. Donc la probabilité d'être inférieur ou égal à 170 sous la distribution $\mathrm{N}(\bar{x}, s^2)$ est la même que la probabilité d'être inférieur ou égal à $(170 - 176.6)/7.99 = -0.83$ sous la distribution $\mathrm{N}(0, 1)$. La probabilité cherchée est la valeur $\Phi(-0.83)$. Cette valeur n'est pas dans le tableau, mais on peut utiliser la symétrie pour obtenir $\Phi(-0.83) = 1 - \Phi(0.83) = 1 - 0.79673 = 0.20327 = 20.327\,\%$. La fréquence relative d'observations inférieures ou égales à 170 cm est 23.3 %, ce qui est proche de la valeur sous le modèle normal.

Exercice 5. (a) La première distribution décrit des données qui sont plus grandes qu'une certaine valeur (par exemple, si cet valeur est 0, les données sont positives), et qui sont en grande partie concentrées près de ce nombre. Le pourcentage attendu d'observations décroit rapidement en fonction de la distance par rapport à ce nombre.

La deuxième distribution décrit des données qui sont contenues entre deux nombres, et qui sont réparties entre eux de façon uniforme.

La troisième distribution décrit des données qui sont plus grandes qu'un certain nombre, et qui sont en grande partie concentrées dans une région à droite de ce nombre. Le pourcentage attendu d'observations décroit en fonction de la distance par rapport à cette région, et ce de façon plus rapide vers la gauche que vers la droite.

La quatrième distribution décrit des données qui sont dispersées de façon symétrique autour d'un certain nombre, et qui sont très concentrées autour de ce nombre. Le pourcentage attendu d'observations décroit rapidement en fonction de la distance par rapport à ce nombre.

(b) La troisième pour l'exercice 2 et la quatrième pour l'exercice 4.

(c) Oui, la quatrième.