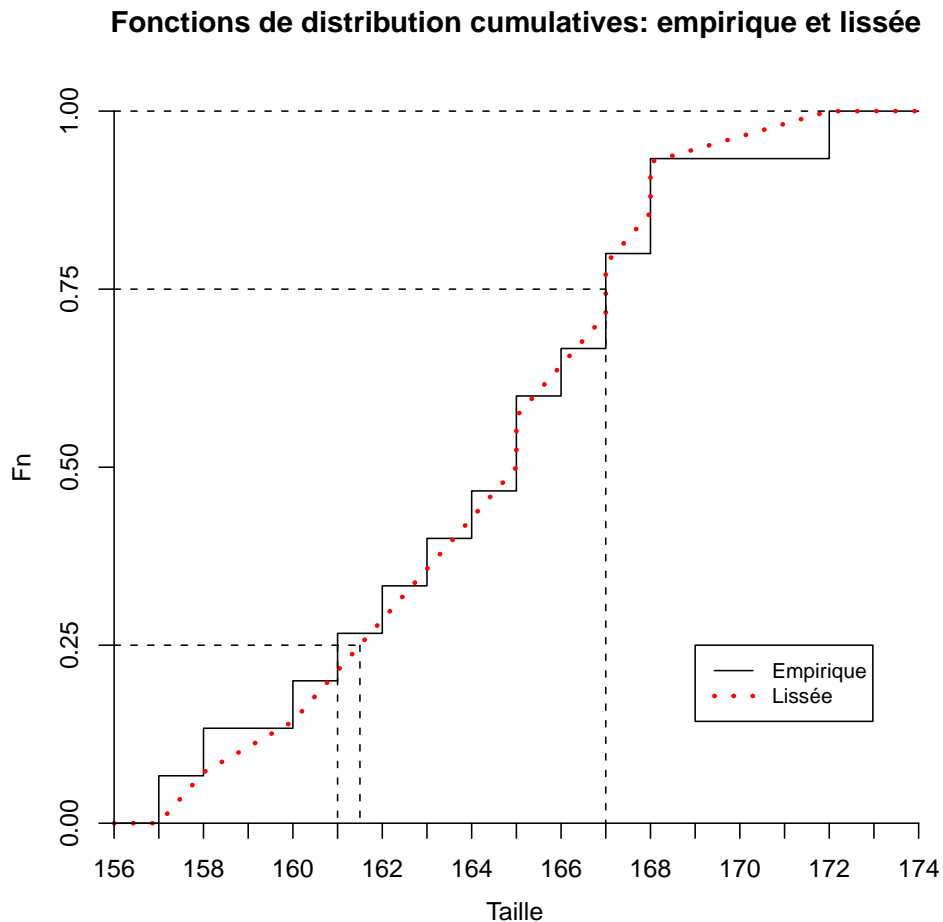


## Travaux pratiques I: réponses

1.1 Si l'on classe les tailles des filles dans l'ordre croissant, on voit que l'observation qui se situe au centre de la distribution (la huitième) vaut 165 cm.

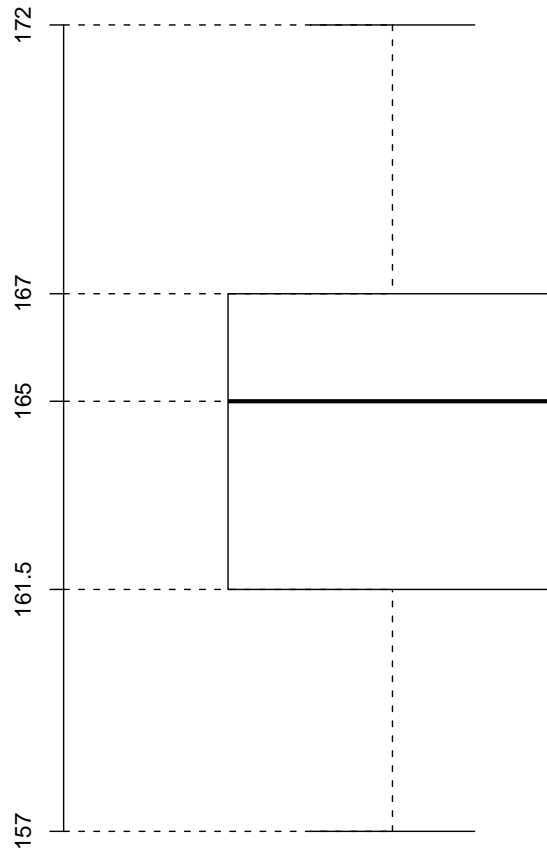
1.2 Pour répondre, on peut se baser soit sur la fonction de distribution cumulative empirique, soit sur une version lissée de celle-ci. Les graphiques de ces fonctions se trouvent dans le polycopié. Il existe plusieurs méthodes de lissage de la cumulative empirique. Pour obtenir une version lissée plus sophistiquée que celle du polycopié, nous avons recours à un logiciel statistique (ici le logiciel R). La méthode par défaut de R nous fournit la courbe en rouge dans la figure ci-dessous.



Soient  $q_{0.25}$  et  $q_{0.75}$  le premier et le troisième quartile des tailles des filles. Avec la cumulative empirique, on obtient  $q_{0.25} = 161$  cm  $q_{0.75} = 167$  cm, ce qui donne un écart interquartile de 6 cm. Avec la version lissée, on obtient  $q_{0.25} = 161.5$  cm  $q_{0.75} = 167$  cm, ce qui donne un écart interquartile de 5.5 cm.

1.3 En utilisant la définition du box-plot se trouvant dans le polycopié et les résultats obtenus ci-dessus avec la version lissée de la fonction de distribution cumulative, on trouve

le graphe ci-dessous:



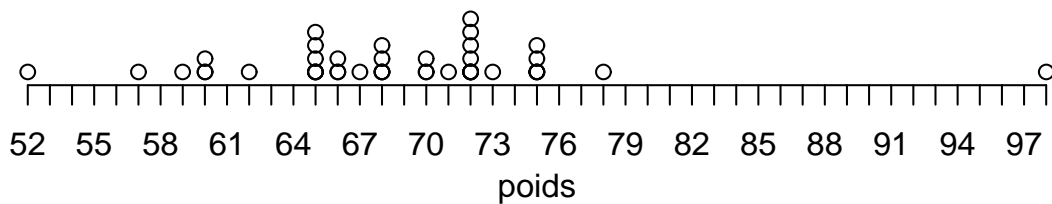
Le boxplot ne signale donc pas d'outliers dans les tailles des filles.

2. En utilisant la formule du polycopié, on obtient  $s \simeq 3.953$ . Ainsi, avec le résultat obtenu dans l'exercice 1. avec la fonction de distribution cumulée lissée, on obtient

$$\frac{I_q}{s} \simeq \frac{5.5}{3.953} \simeq 1.391.$$

On est donc assez près de la valeur théorique pour une distribution normale (1.349); l'écart relatif est inférieur à 5%.

3.1 On trouve la répartition suivante:



3.2 On trouve 68.6 kg pour la moyenne et 68 kg pour la médiane.

3.3 Après suppression de la donnée la plus élevée, on trouve 67.6 kg pour la moyenne et toujours 68 kg pour la médiane.

3.4 On voit que la moyenne change de 1 kg (environ 1.5% de sa valeur) et que la médiane ne bouge pas.

3.5 Si l'on souhaite une mesure qui caractérise la position de "la majorité" des individus, il vaut mieux utiliser la médiane, car comme on l'a vu il suffit d'un seul cas extrême pour fausser sensiblement la moyenne. De plus, la connaissance de la médiane d'une population permet de situer son propre poids (ou celui d'un autre sujet) dans la moitié la plus lourde ou dans la moitié la plus légère de la population. Cela dit, on utilise généralement la moyenne lorsqu'on s'intéresse à un "total" (exemple: pour le département des finances, le salaire moyen pourrait être plus utile que le salaire médian, car il permet d'estimer le bénéfice attendu des impôts). Par contre, pour le problème présent, il est plus probable que l'on cherche à situer un autre poids par rapport au "poids de la plupart des gens" plutôt que de s'intéresser au poids total de la population. Il faut encore signaler une autre raison qui fait que la moyenne est souvent préférée à la médiane: elle a des propriétés mathématiques appréciables, comme la linéarité. Mais ici, nous menons une analyse exploratoire, sans calculs particuliers, et donc la médiane est préférable.

4.1 A l'aide de la courbe de croissance correspondante, on trouve que la taille médiane des filles de 18 ans est d'environ 165 cm.

4.2 Comme la distribution semble légèrement asymétrique à gauche (i.e. plus étalée dans la partie inférieure de la courbe) on peut présumer que la moyenne est légèrement inférieure à la médiane (peut-être autour de 164cm).

Les tailles étant notoirement limitées (on peut quasiment exclure l'existence de cas de moins de 50cm ou de plus de 250cm), il serait difficile dans le cas présent que la moyenne soit très différente de 164 avec les courbes de croissance fournies.

Cela dit, si on ne connaissait absolument rien de la population en question (imaginons que les courbes correspondent aux tailles d'extra-terrestres), on pourrait avoir des surprises. Si par exemple 3% de la population avait une taille de 30m, les courbes ne changeraient pas mais la moyenne pourrait être très différente de 164cm.

4.3 A l'aide de la même courbe de croissance, on trouve que l'écart interquartile de tailles des filles de 18 ans est d'environ 8 cm.

4.4 Comparons les quantiles:

- Pour la population:

$qp_{0.1} \simeq 157$  cm,  $qp_{0.25} \simeq 161.5$  cm,  $qp_{0.5} \simeq 165$  cm,  $qp_{0.75} \simeq 170$  cm,  $qp_{0.9} \simeq 172.5$  cm.

- Pour l'échantillon du polycopié (en utilisant la cumulative empirique):

$qe_{0.1} = 158$  cm,  $qe_{0.25} = 161$  cm,  $qe_{0.5} = 165$  cm,  $qe_{0.75} = 167$  cm,  $qe_{0.9} = 168$  cm.

On voit que les quantiles sont très similaires. Les différences légèrement plus importantes dans la partie supérieure de la distribution ne sont probablement pas significatives, étant donnée la petite taille de l'échantillon. On ne remarque donc pas de différence importante entre les deux distributions.

4.5 Les petites différences sont probablement dues principalement à la petite taille de l'échantillon.

4.6 D'après la courbe de croissance correspondante, le poids de 42 kg se situe entre la médiane et le troisième quartile des poids des filles de 12 ans. La probabilité que le poids d'une fille de 12 ans prise au hasard dépasse 42 kg est donc comprise entre 50% et 25%.

4.7 L'observation de la courbe de croissance correspondante nous révèle que la distribution du BMI des filles de 16 ans est assez asymétrique. Le modèle de Gauss n'est donc pas approprié.

4.8 Contrairement à l'exemple précédent, la distribution du périmètre crânien des garçons de 17 ans semble assez symétrique. Le modèle de Gauss n'est donc pas exclu d'emblée, mais il conviendrait d'analyser plus précisément la situation à l'aide d'outils statistiques qui dépassent le cadre de ce cours avant de l'adopter.

4.9 Pour une fille de 15 ans, un BMI de  $15 \text{ kg/m}^2$  est largement au dessous du quantile 3% de la distribution correspondante dans la population. Cela signifie que moins de 3% des filles de cet âge ont un BMI égal ou inférieur à celui-ci. On peut donc dire que cette valeur n'est pas dans la norme.

4.10 Dans la table du photocopié, on trouve que le garçon de 98 kg mesure 179 cm. son BMI vaut donc  $98/1.79^2 \text{ kg/m}^2 = 30.59 \text{ kg/m}^2$ . En considérant la courbe de croissance correspondante, on constate que cette valeur est largement au dessus du quantile 97% du BMI des garçons de 19 ans dans la population. Cette valeur n'est donc pas dans la norme.

## 5. Observations:

- La distribution du BMI est, de manière générale, beaucoup plus asymétrique aux Etats-Unis qu'en Suisse.
- En Suisse, l'asymétrie à droite (i.e. étalement de la partie supérieure de la distribution) est de plus en plus marquée jusqu'à l'âge de 19 ans environ, âge auquel a lieu une stabilisation, voire même une réduction chez les filles. Aux Etats-Unis, la stabilisation n'a pas lieu.
- La médiane de la distribution est de façon générale plus élevée aux Etats-Unis qu'en Suisse.
- Dans les deux pays, les filles et les garçons ont des valeurs similaires du BMI médian à tous les âges, avec un léger dépassement du BMI médian des garçons vers 20 ans. Par contre, aux Etats-Unis c'est chez les filles que l'on observe le plus de personnes au BMI très élevé, alors qu'en Suisse c'est chez les garçons (la différence étant moins marquée en Suisse).
- Les différences entre les deux pays sont situées principalement dans la partie droite de la distribution du BMI (i.e. le domaine des BMI supérieurs à la médiane).

6.1 et 6.2 Le graphique d'une fonction de distribution cumulative représente le pourcentage d'observations (mesures) d'une certaine variable qui sont inférieures ou égales à chacune des valeurs possibles de cette variable. Le diagramme de la Figure 12-16 représente le pourcentage de "open channels" en fonction de la variable "membrane potential". A première vue, ce diagramme n'est donc pas celui d'une fonction de distribution cumulative, mais plutôt celui d'une probabilité conditionnelle d'ouverture.

(Toutefois, comme le diagramme semble croissant, on peut postuler que pour chaque canal il existe un potentiel minimal au-dessus duquel le canal reste ouvert. Dans ce cas, on peut considérer que le diagramme représente la fonction de distribution cumulative de la variable "potentiel minimal d'ouverture du canal ionique".)

6.3 et 6.4 La forme globale de l'histogramme n'est pas gaussienne car elle comporte deux pics (le premier se situe en zéro). Toutefois, on peut imaginer que les données représentées dans cet histogramme sont un mélange de deux types de données: d'une part une mesure de courant avec distribution gaussienne (qui produit la deuxième bosse) et d'autre part une autre variable avec beaucoup de valeurs proches de zéro (qui produit la première bosse). Cette autre variable est en fait le bruit de l'enregistrement du patch-clamp.

6.5 La courbe est un lissage de l'histogramme. En la multipliant par un nombre tel que la surface sous la courbe devienne égale à 1, on obtiendrait une estimation de la densité de la variable représentée.

6.6 Vraisemblablement pas, car le pic en zéro fait diminuer la moyenne. Toutefois, si on s'intéresse uniquement aux données produisant la deuxième bosse, alors la valeur du pic donne une bonne idée de leur moyenne.

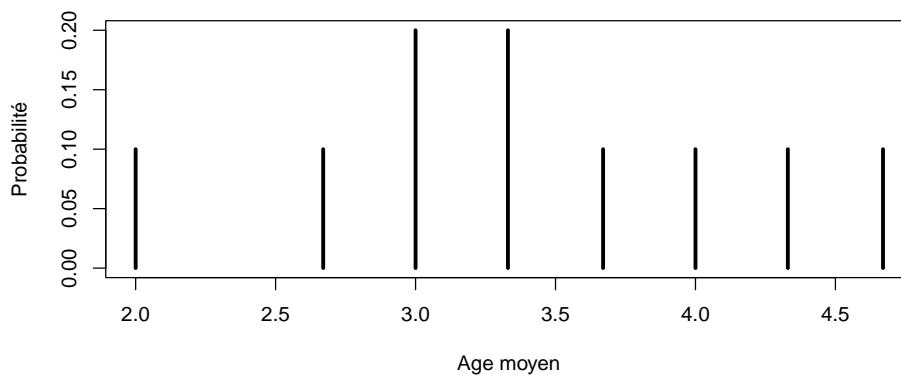
7. On a 10 échantillons possibles:

$\{1, 2, 3\}, \{1, 2, 5\}, \{1, 2, 6\}, \{1, 3, 5\}, \{1, 3, 6\}, \{1, 5, 6\}, \{2, 3, 5\}, \{2, 3, 6\}, \{2, 5, 6\}, \{3, 5, 6\}$ .

Les moyennes correspondantes sont respectivement:

2, 2.67, 3, 3, 3.33, 4, 3.33, 3.67, 4.33, 4.67.

Tous les échantillons sont équiprobables et sont tirés avec une probabilité de 1/10. On obtient donc la distribution de probabilité suivante pour l'âge moyen des individus d'un échantillon de taille 3:



8. Soit  $m(X)$  la moyenne des  $x_1, \dots, x_n$ .  $s^2(X)$  est la moyenne des carrés des écarts entre  $x_1, \dots, x_n$  et  $m(X)$ . Appelons ces écarts  $e_1, \dots, e_n$  ( $e_1 = x_1 - m(X), \dots, e_n = x_n - m(X)$ ).

- Si on multiplie les  $x_1, \dots, x_n$  par 3, tous les écarts sont aussi multipliés par 3. Donc tous les écarts au carré sont multipliés par 9, de même que leur moyenne. Donc  $s^2(X)$  est multiplié par 9. Ainsi  $s(X) (= \sqrt{s^2(X)})$  est multiplié par 3 ( $= \sqrt{9}$ ). On obtient donc  $s(3X) = 3s(X) = 6$ .
- Si on ajoute 4 à chacun des  $x_1, \dots, x_n$ , les écarts  $e_1, \dots, e_n$  ne changent pas, tout est simplement décalé vers la droite. Ainsi,  $s(X)$  ne change pas non plus et on obtient  $s(X + 4) = s(X) = 2$ .
- On a vu que si on multiplie une variable par un nombre son écart-type est multiplié d'autant, et que si on lui ajoute quelque chose son écart-type ne change pas. En combinant ces deux propriétés on arrive facilement à la conclusion que  $s(3X + 4) = s(3X) = 3s(X) = 6$ .