Étude de l'échantillonnage
Partie 2 : Intervalles de confiance

Feuille de travail de l'élève

1. Distribution normale

Examinez de nouveau les données tirées des 100 échantillons différents de 20 tailles (voir l'annexe A).

Rappel :

La moyenne des moyennes des échantillons est approximativement égale à la moyenne de la population $\mu \cong 164,5$	**L'écart-type des moyennes des échantillons est à peu près égal à l'erreur-type.** $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10,2}{\sqrt{20}} = 2,28$

De toute évidence, la distribution des moyennes des échantillons semble être normale. (En fait, si l'on sélectionnait tous les échantillons possibles de 20 tailles, la distribution serait parfaitement normale.) À partir de ce résultat, répondez aux questions suivantes :

a) Quel pourcentage des moyennes des échantillons devrait être compris dans l'intervalle de plus ou moins une erreur-type par rapport à la moyenne des moyennes des échantillons?
b) Quel pourcentage devrait être compris dans l'intervalle de plus ou moins deux erreurs-types?
c) Vérifiez ces prévisions en utilisant les moyennes des échantillons fournies à l'annexe A. La situation est-elle exactement celle à laquelle vous vous attendiez? Expliquez.

Que pouvez conclure à propos d'un seul échantillon aléatoire de 20 tailles?

On pourrait conclure que si nous sélectionnions **un** échantillon de 20 élèves au hasard à partir de cette population, la moyenne de l'échantillon aurait 68 % de probabilités d'être comprise dans l'intervalle de plus ou moins une erreur-type par rapport à la moyenne de la population, et 95 % de probabilités d'être comprise dans l'intervalle de plus ou moins deux erreurs-types par rapport à la moyenne de la population.

2. Théorème central limite

Nous avons vu que la taille de l'échantillon joue un rôle dans la distribution des moyennes des échantillons. Pour nous assurer que les moyennes des échantillons aient une distribution normale, nous avons choisi des échantillons de 30 élèves et plus. Ainsi, nous pouvons énoncer le théorème central limite suivant :

Si $n \geq 30$**, la distribution des moyennes,** $\overline{X}$**, de tous les échantillons aléatoires de taille** n **est approximativement normale, de moyenne** μ **et d'écart-type** $\frac{\sigma}{\sqrt{n}}$**.**

Examinez le scénario suivant :

Imaginez que vous essayez de déterminer la taille moyenne des élèves de votre école. Comme il n'est guère pratique de déterminer la taille de tous les élèves, vous recueillez des données sur les tailles auprès d'un échantillon aléatoire de 30 élèves.

- À votre avis, la moyenne de cet échantillon sera-t-elle exactement la même que celle de la population?
- Selon vous, la moyenne des échantillons s'approchera-t-elle de la moyenne de la population? Si oui, jusqu'à quel point ces moyennes seront-elles proches?

On dit que la taille moyenne calculée pour votre échantillon aléatoire est une *estimation ponctuelle*. Il s'agit d'une variable statistique unique utilisée pour estimer la moyenne de la population. L'expérience nous dit que si nous choisissions un autre échantillon aléatoire, nous obtiendrions probablement une moyenne différente et, donc, une estimation ponctuelle différente, et ce, en raison de la variation des moyennes des échantillons.

Pour tenir compte de cette variation des moyennes, nous déterminons un intervalle dans lequel nous estimons que la moyenne réelle de la population se trouvera en considérant la distribution de la moyenne de chaque échantillon. Selon le degré de confiance que nous voulons avoir que l'intervalle contiendra la moyenne de la population, nous pouvons choisir des intervalles de tailles différentes. Plus l'intervalle est grand, plus nous pouvons avoir confiance qu'il contienne la valeur en question. Y a-t-il un inconvénient à choisir un grand intervalle? En général, nous utilisons des intervalles nous permettant de dire avec 95 % de confiance qu'ils contiennent la valeur.

Un intervalle de confiance de 95 % signifie que, si nous sélectionnions tous les échantillons possibles de même taille, dans 95 % des échantillons, la moyenne réelle de la population serait incluse dans l'intervalle de confiance autour de la moyenne de chaque échantillon.

Exercice : Analysez divers échantillons

Examinez les quatre diagrammes figurant à la page suivante. Les échantillons sont sélectionnés à partir de l'ensemble des données originales sur les tailles, avec $\mu = 164,7$ et $\sigma = 10,21$. L'échantillon n^{o} 1 est défini ci-dessous.

Échantillon n^{o} 1

Données : $n = 30$ et $\overline{X} = 165,9$

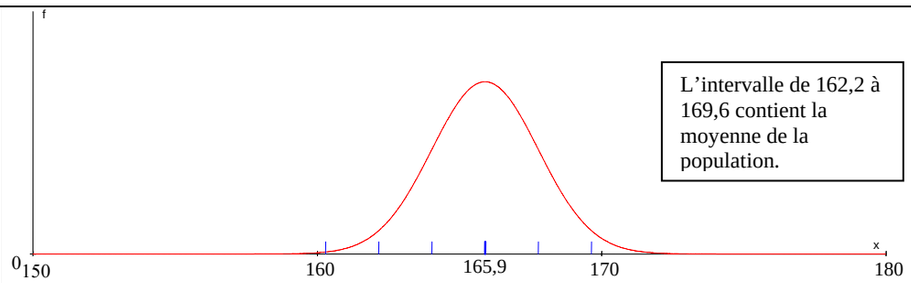
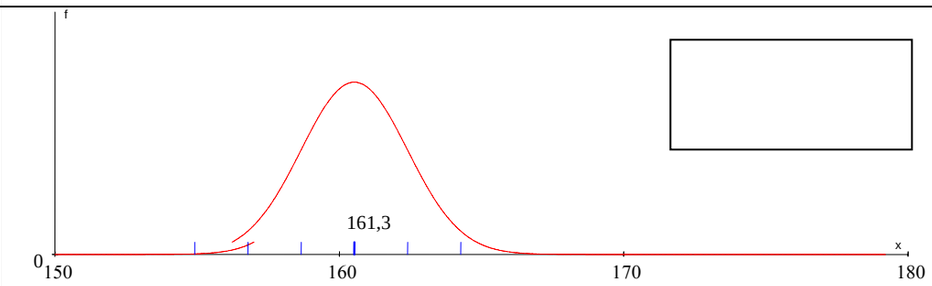
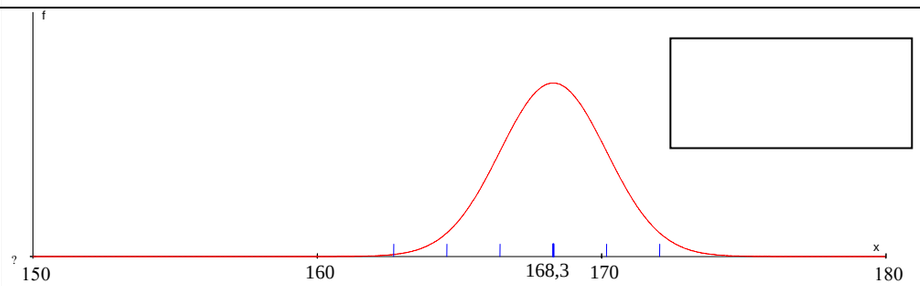
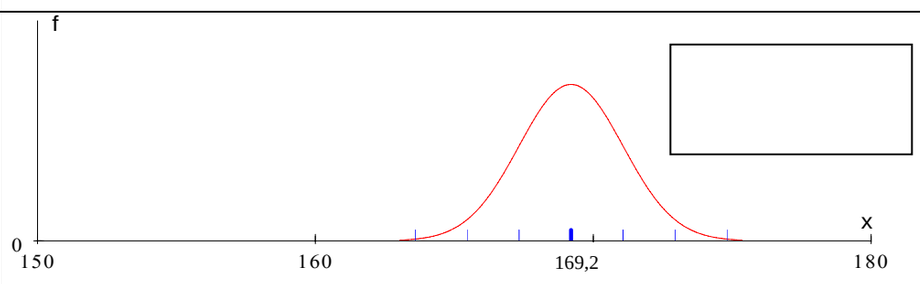
Donc, $\sigma_{\overline{X}} = \dfrac{10,2}{\sqrt{30}} = 1,86.$

Déterminez l'intervalle de plus ou moins $2\sigma_{\overline{X}}$ par rapport à $\overline{X}$.

$$\overline{X} - 2\sigma_{\overline{X}} = 165,9 - 2(1,86) = 162,2$$
$$\overline{X} + 2\sigma_{\overline{X}} = 165,9 + 2(1,86) = 169,6$$

Contient-il la moyenne de la population?

Faites des calculs comparables pour les échantillons n^{os} 2, 3 et 4. Pour chaque échantillon, déterminez si l'intervalle contient la moyenne de la population.

Échantillon n^o 1 $n = 30$ $\overline{X} = 165{,}9$ $\sigma_{\overline{X}} = \frac{10{,}2}{\sqrt{30}} = 1{,}86$	
Échantillon n^o 2 $n = 30$ $\overline{X} = 161{,}3$ $\sigma_{\overline{X}} = \frac{10{,}2}{\sqrt{30}} = 1{,}86$	
Échantillon n^o 3 $n = 30$ $\overline{X} = 168{,}3$ $\sigma_{\overline{X}} = \frac{10{,}2}{\sqrt{30}} = 1{,}86$	
Échantillon n^o 4 $n = 30$ $\overline{X} = 169{,}2$ $\sigma_{\overline{X}} = \frac{10{,}2}{\sqrt{30}} = 1{,}86$	

Comme la distribution est normale, nous pouvons supposer que toute moyenne d'échantillon a 95 % de probabilités d'être dans l'intervalle de plus ou moins deux $\sigma_{\overline{X}}$ par rapport à la moyenne de la population.

De toute évidence, les échantillons n'auront pas tous une moyenne se situant à plus ou moins deux $\sigma_{\overline{X}}$ de μ. Toutefois, nous nous attendons à ce que 95 % des moyennes des échantillons soient situées à plus ou moins deux $\sigma_{\overline{X}}$ de μ.

Si nous connaissons l'écart-type de la population, nous pouvons utiliser l'erreur-type pour déterminer l'intervalle autour de la moyenne de chaque échantillon pour lequel il existe 95 % de probabilités qu'il contienne la moyenne de la population.

Dans la réalité, nous ne connaissons généralement pas la moyenne et l'écart-type de la population. Souvenez-vous du scénario où nous voulions déterminer la taille moyenne des élèves de votre école. Dans ce cas, nous connaissions seulement les tailles d'un échantillon aléatoire de 30 élèves.

Que se passe-t-il si vous ne connaissez pas σ ?

Vous devez à nouveau examiner l'information dont nous disposons et la façon dont elle est liée à σ.

Examinez les écarts-types des échantillons de l'exercice. Pour commencer, vous devez savoir que le calcul de l'écart-type d'un échantillon est légèrement différent de celui de l'écart-type de la population. ($s_X = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$). Notez que le menu 1-Var Stats de votre calculatrice TI-83 donne à la fois les fonctions S_X (écart-type de l'échantillon) et σ_X (écart-type de la population).

La différence est qu'au lieu de diviser par la taille de l'échantillon (*n*), vous diviserez par *n-1*. Pour calculer s_X, vous devez d'abord connaître quelle est $\overline{X}$. Par conséquent, *n-1* valeurs seulement de l'échantillon sont libres de varier. La n^e valeur est établie, puisque $\overline{X}$ a également été déterminé.

Par exemple, supposons que la moyenne d'un échantillon de trois valeurs est égale à 10. Comme la moyenne est de 10 et que $n = 3$, il ne faut connaître que deux valeurs distinctes pour que la troisième soit fixée; elle est déterminée par l'information, puisque la somme doit être égale à 30.

Examinez les valeurs de s_X à l'annexe A. Elles sont assez près de la valeur de σ. En fait, un statisticien nommé William S. Gosset a élaboré la distribution appelée *loi de Student*. Selon ses travaux, il est raisonnable de remplacer σ par s_X dans notre formule de $\sigma_{\overline{X}}$, ce qui nous donne $\sigma_{\overline{X}} = \frac{s_X}{\sqrt{n}}$.

Donc, l'intervalle de confiance à 95 % de la moyenne de la population est donné approximativement par la formule suivante :

$\overline{X} \pm 2\frac{s_X}{\sqrt{n}}$, qui signifie que $\overline{X} - 2\frac{s_X}{\sqrt{n}} \le \mu \le \overline{X} + 2\frac{s_X}{\sqrt{n}}$

Les variables peuvent être déterminées à partir de l'échantillon aléatoire, si $n \ge 30$. En outre, nous pouvons énoncer un intervalle pour la population estimée et déterminer un niveau de confiance.

3. Projet

Allez sur le site Web *Recensement à l'école* à www.censusatschool.ca ou sur tout autre site où vous pouvez obtenir des données fiables. Choisissez quelle information quantitative vous aimeriez examiner et quelle population vous souhaiteriez échantillonner. Faites les calculs et rédigez un bref rapport — qui prendra la forme d'un article de journal — sur cette population en vous inspirant des résultats de votre échantillon.

Collaboration : Anna Spanik, professeure de mathématiques, école secondaire Halifax West, Nouvelle-Écosse.