

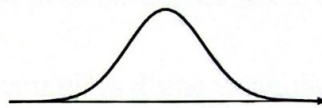
5 Statistiques inférentielles

L'inférence statistique a pour but de déterminer les caractéristiques d'une **population** en utilisant celles d'un **échantillon** de cette population.

Puisque les données utilisées ne sont pas celles de **toute** la population, il n'est pas possible de déterminer de manière exacte les caractéristiques de celle-ci. Néanmoins, la statistique inférentielle permet d'estimer ces caractéristiques, tout en fournissant une information sur la précision des mesures obtenues.

5.1 La loi normale

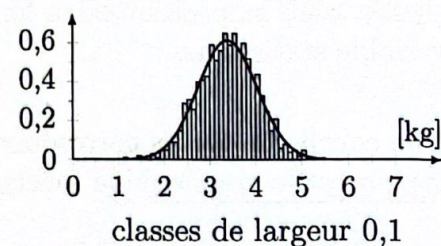
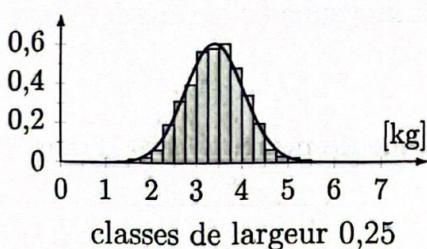
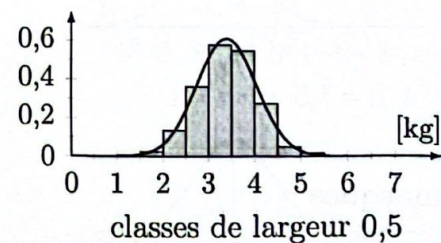
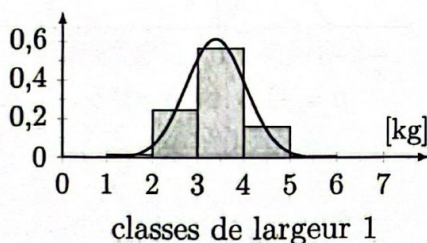
Dans de nombreux contextes, une variable statistique⁴ se distribue selon une **courbe en cloche**, aussi appelée **courbe de Gauss**, dont l'allure est la suivante :



Cette courbe en cloche correspond à la courbe de fréquence **théorique** de la variable statistique. Si on disposait d'un échantillon extrêmement grand, et que les données étaient regroupées en classes très petites, le polygone des fréquences ressemblerait à cette courbe.

Exemple

Allures des histogrammes du poids de 1000 nourrissons à la naissance.



4. De nombreux raccourcis théoriques ont dû être pris afin d'intégrer la matière dans le plan d'étude de l'École de Culture Générale. En particulier, la notion essentielle de variable aléatoire est escamotée au bénéfice de celle de variable statistique.

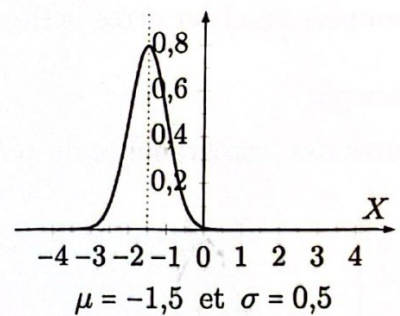
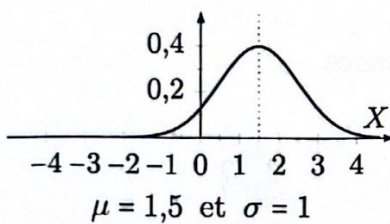
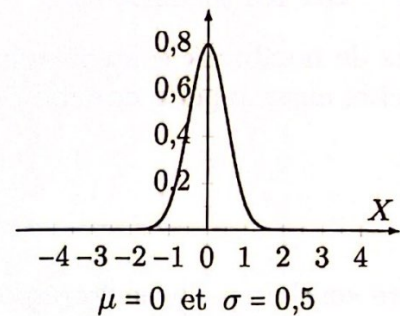
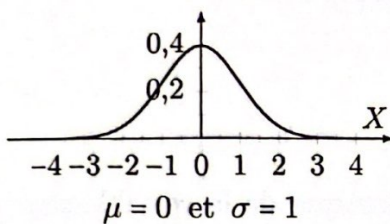
Lorsque la fréquence théorique d'une variable statistique X se distribue selon une courbe de Gauss, on dira que la variable suit une loi normale. Si X possède une moyenne μ et une variance σ^2 , on écrira

$$X \sim \mathcal{N}(\mu; \sigma^2)$$

et on dira que X suit une loi normale de moyenne μ et de variance σ^2 .

Exemple

Courbes de Gauss correspondant à différentes moyennes et écarts-types.



Remarques

1. La courbe de distribution d'une variable suivant une loi normale ressemble toujours à une cloche, mais sa position et sa forme sont déterminées par la moyenne et la variance de la variable statistique.
2. Une courbe de Gauss correspondant à une variable statistique de moyenne μ et d'écart-type σ est le graphe de la fonction f donnée par

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

3. L'aire totale sous une courbe de Gauss vaut toujours 1 et la courbe de Gauss est symétrique par rapport à la moyenne μ .

5.1.1 La loi normale centrée réduite

Une variable statistique qui suit une loi normale de moyenne $\mu = 0$ et de variance $\sigma^2 = 1$, est appelée **variable normale centrée réduite**. Dans ce cas, la variable sera souvent notée Z , et on aura

$$Z \sim \mathcal{N}(0; 1)$$

Les fréquences cumulées croissantes d'une variable statistique nous permettaient de répondre à des questions du type :

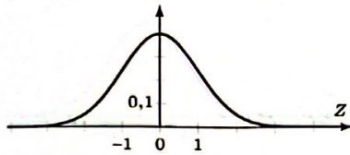
Quelle est la proportion des poissons pêchés lors d'un concours dont la longueur est plus petite que 30 cm ?

Cette même question pourrait être reformulée en termes de probabilités :

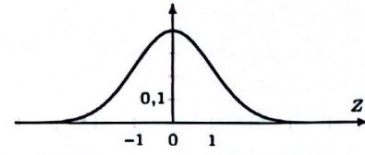
Quelle est la probabilité qu'un poisson tiré au hasard parmi les prises du concours mesure moins de 30 cm ?

En d'autres termes, les fréquences cumulées d'une variable statistique permettent de déterminer certaines probabilités. Dans le cas des variables statistiques suivant une loi normale, la courbe de Gauss se révèle plus flexible que les fréquences cumulées.

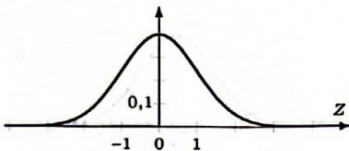
Une aire sous la courbe de Gauss représente une probabilité.



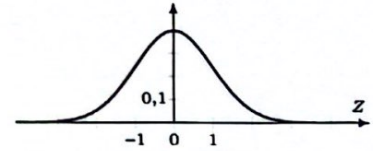
$$Z = \text{nbre} \text{ et } P(Z = \text{nbre}) =$$



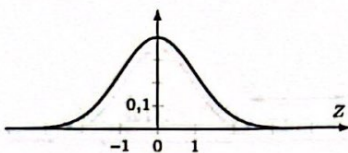
$$Z \in \mathbb{R} \text{ et } P(Z \in \mathbb{R}) =$$



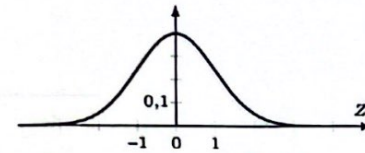
$$Z \leq 0 \text{ et } P(Z \leq 0) =$$



$$Z \leq \text{nbre} \text{ et } P(Z \leq \text{nbre}) =$$



$$P(Z \leq \text{nbre}) = P(Z < \text{nbre})$$



$$P(Z > \text{nbre}) = 1 - P(Z \leq \text{nbre})$$

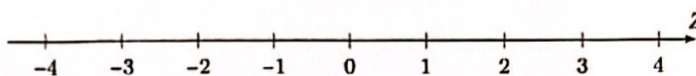
Rappel

La **valeur absolue** d'un nombre x , notée $|x|$, est ce nombre "sans son signe". Plus précisément : Si x est positif, alors $|x| = x$; si x est négatif, alors $|x| = -x$. Par exemple,

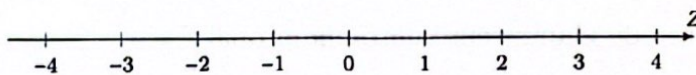
$$\begin{array}{lll} |3| = & |-3| = & |12,5| = \\ |-\sqrt{2}| = & |0| = & |-1/7| = \end{array}$$

Exemple

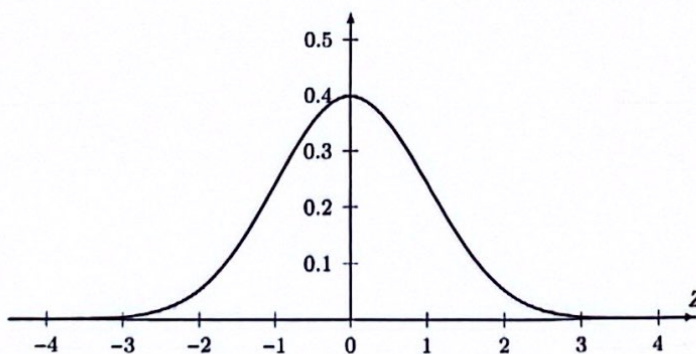
- Représenter sur l'axe Z toutes les valeurs vérifiant $|Z| \leq 2,5$.



- Représenter sur l'axe Z toutes les valeurs vérifiant $|Z| > 1,5$.



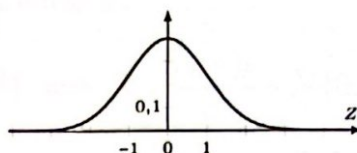
- Pour $Z \sim \mathcal{N}(0; 1)$, représentez graphiquement $P(|Z| > 1,5)$.



5.1.2 Calculs avec la table de la loi normale

La table numérique pour la loi normale centrée réduite à la page 129 donne la probabilité d'obtenir pour une variable $Z \sim \mathcal{N}(0;1)$ une valeur inférieure ou égale à une valeur fixée $a \geq 0$.

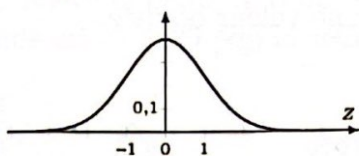
$P(Z \leq a)$:



D'autres probabilités peuvent se calculer en exploitant la symétrie de la courbe de Gauss, ainsi que le fait que l'aire totale sous la courbe vaut 1.

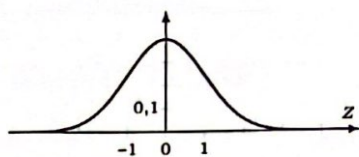
Exemples

1.



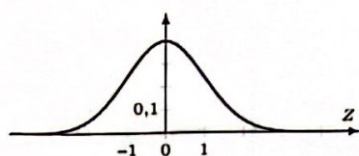
$$P(Z \leq 1,35)$$

2.



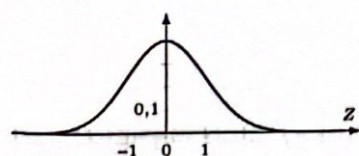
$$P(Z > 1,35)$$

3.



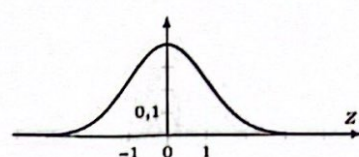
$$P(Z < -0,5)$$

4.



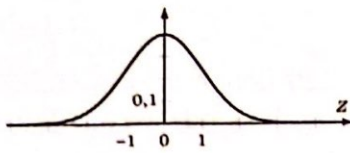
$$P(Z > -0,5)$$

5.



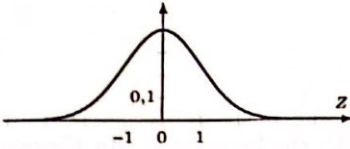
$$P(1 < Z < 2)$$

6.



$$P(-1,34 < Z < 2,11)$$

7.

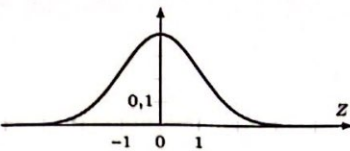


$$P(|Z| > 1,65)$$

On peut également utiliser la table “dans l’autre sens” si on cherche une valeur a pour que la probabilité que Z soit supérieure ou inférieure à a soit égale à une valeur donnée.

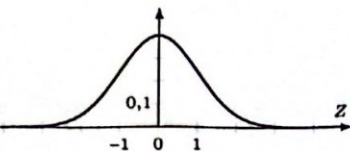
Exemples

1.



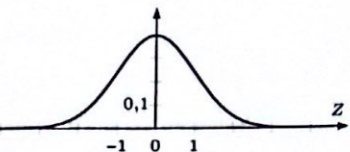
Déterminez a tel que $P(Z \leq a) = 75\%$

2.



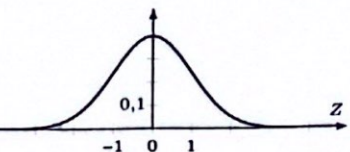
Déterminez b tel que $P(Z \geq b) = 10\%$

3.



Déterminez c tel que $P(|Z| < c) = 80\%$

4.



Déterminez d tel que $P(|Z| \geq d) = 5\%$

5.1.3 Normalisation et cote Z

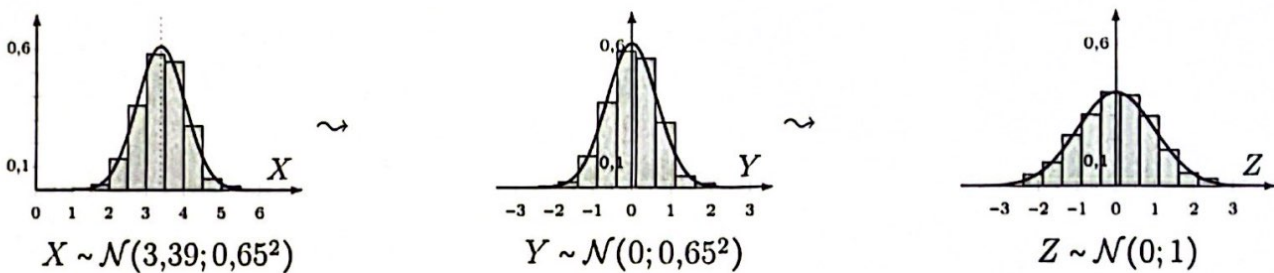
Si une variable statistique X suit une loi normale de moyenne μ et de variance σ^2 , alors, en soustrayant μ de toutes les modalités de X , on obtient une nouvelle variable normale $Y = X - \mu$ de moyenne 0 et de variance σ^2 . En divisant les modalités de Y par σ , on obtient une variable Z de moyenne 0 et de variance 1 :

$$X \sim \mathcal{N}(\mu; \sigma^2) \implies Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0; 1)$$

La variable normale centrée réduite Z est appelée la **cote Z** de X , obtenue par **normalisation**.

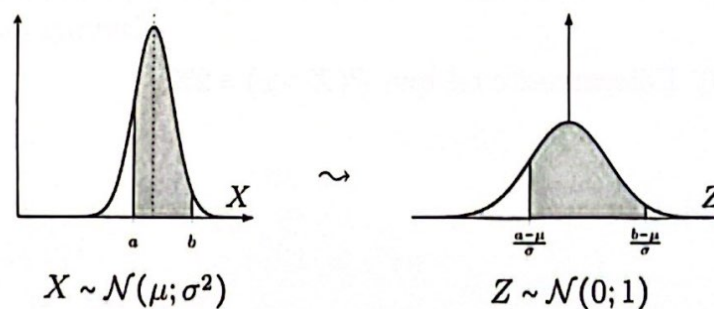
Exemple

On relève le poids X d'un échantillon de 1000 nourrissons à leur naissance ; la moyenne des poids est de 3,39 [kg] et leur écart-type est de 0,65.



Remarque

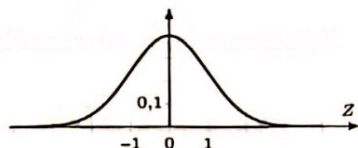
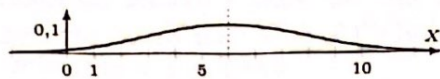
La normalisation d'une courbe de Gauss préserve les aires sous la courbe :



Cette remarque nous permet de calculer les probabilités correspondant à n'importe quelle variable statistique $X \sim \mathcal{N}(\mu; \sigma^2)$ à l'aide de sa cote Z .

Exemple

1. Soit $X \sim \mathcal{N}(6; 9)$. Déterminez $P(X \leq 4)$.



2. Soit $X \sim \mathcal{N}(-1; 0,01)$. Calculez $P(X > 0)$.

Comme dans le cas d'une loi normale centrée réduite, étudions un exemple dans lequel on cherche une valeur a pour que la probabilité que X soit supérieure ou inférieure à une valeur donnée.

Exemple

Soit $X \sim \mathcal{N}(3000; 40000)$. Déterminez c tel que $P(X > c) = 2\%$.

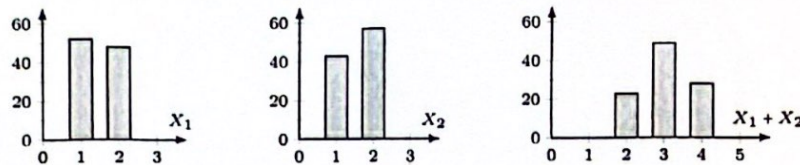
Pour terminer ce paragraphe, traitons un exemple où la théorie qui précède est utilisée dans une application plus concrète.

5.2 Théorème Central Limite (TCL)

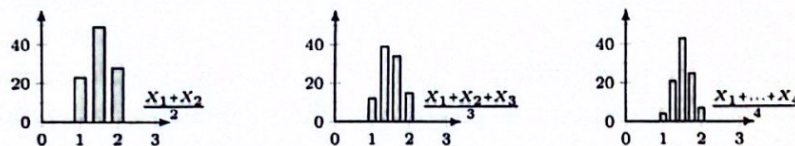
Commençons par un exemple.

Trente employés d'une entreprise vont régulièrement à leur pause de 10h prendre un café à la cafétéria. Des cafés en poudre à 1 CHF et des cafés machines à 2 CHF sont disponibles. La gérante de la cafétéria a observé qu'aucun des employés n'avait de préférence particulière dans leur choix de café. Elle en déduit que chaque employé paiera **en moyenne** environ 1,50 CHF par jour pour ce café du matin. Mais pour prévoir ses coûts et planifier ses dépenses, la gérante aurait surtout besoin de savoir, avec une certitude de 95% par exemple, quels seraient ses revenus minimaux et maximaux par jour.

Notons X_1, X_2, \dots, X_{30} les variables statistiques représentant le type de café choisi par chacun des 30 employés pendant les 100 premiers jours de l'année. Voici par exemples les histogrammes de X_1 , de X_2 , ainsi que de la variable statistique $X_1 + X_2$ qui représente le total journalier des dépenses des deux premiers employés.



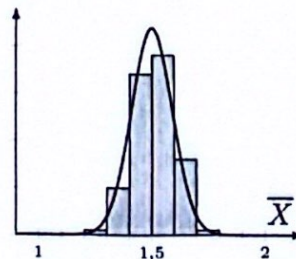
Plutôt que le total $X_1 + X_2$, représentons la **moyenne** des dépenses $\frac{X_1 + X_2}{2}$, et procédons de même pour $\frac{X_1 + X_2 + X_3}{3}$ et $\frac{X_1 + \dots + X_4}{4}$:



En continuant ce processus pour les 30 employés, nous obtenons des dépenses journalières entre 30 CHF et 60 CHF. Il devient alors judicieux de regrouper les données en classes. L'historgramme de la variable statistique

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_{30}}{30},$$

qui représente la moyenne du total des dépenses journalières des 30 employés de l'entreprise, a l'allure suivante.



La moyenne \bar{X} des variables statistiques X_1, X_2, \dots, X_{30} semble suivre une loi normale ! Ici, $\bar{X} \sim \mathcal{N}(1,5; 1/120)$.

En termes de probabilités, l'exemple précédent s'interprète comme suit. Si on choisit au hasard **un** des cents jours recensés pour X_1 , le prix x_1 du café choisi ce jour-là aura environ 50% de chances d'être de 1 CHF et 50% de chances d'être de 2 CHF. De même, en choisissant au hasard un jour pour chacun des 29 autres employés, on obtiendra des prix x_2, x_3, \dots, x_{30} de 1 CHF ou 2 CHF à chaque choix.

La moyenne de ces prix $\bar{x} = \frac{x_1 + x_2 + \dots + x_{30}}{30}$ prendra une valeur entre 1 CHF et 2 CHF avec une probabilité donnée par la loi normale. Par exemple, la probabilité que le prix moyen soit moins de 1,70 CHF sera

$$P(\bar{X} \leq 1,70) \cong P(Z \leq 2,19) \cong 98,57\%.$$

Théorème (TCL)

Soient X_1, X_2, \dots, X_n des variables statistiques de même moyenne μ et écart-type σ , et suivant la même courbe des fréquences théorique. Si

— n est suffisamment grand⁵, **ou**

— chaque X_i suit une loi normale, $X_i \sim \mathcal{N}(\mu; \sigma^2)$ pour $i = 1, \dots, n$,

alors le TCL s'applique, et la variable statistique moyenne

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

s'approche d'une loi normale de moyenne μ et d'écart-type $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ ⁶ et on écrit :

$$\bar{X} \sim \mathcal{N}(\mu; \sigma^2/n).$$

Remarques

1. Il ne faut pas confondre l'écart-type de la population, noté σ , et l'écart-type de \bar{X} , noté $\sigma_{\bar{X}}$.
2. Afin de pouvoir appliquer le TCL, nous conviendrons que si X est une variable statistique décrivant une certaine caractéristique d'une population, alors tout échantillon de la population aura cette caractéristique décrite par une variable statistique X_1, X_2, \dots de même type, en particulier ayant même moyenne et écart-type.

5. Une étude plus approfondie de la variable statistique moyenne \bar{X} est nécessaire pour déterminer si n est assez grand. Parfois $n = 20$ ou $n = 30$ suffit, mais d'autres fois, il faut $n = 100$, voire **beaucoup** plus. Dans ces notes, nous dirons que n est "suffisamment grand" pour dire qu'une telle étude des données a montré que le n proposé convient.

6. Si l'échantillon est grand par rapport à la population, c'est-à-dire si $n > \frac{N}{20}$ (où n est la taille de l'échantillon et N celle de la population), il est souvent d'usage de corriger l'écart-type de \bar{X} par un **facteur de correction**. Cet écart-type devient alors

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}.$$

Exemple

Les données historiques d'un certain supermarché montrent que la variable statistique X représentant les dépenses par cliente un 24 décembre a une moyenne de 250 CHF et un écart-type de 70 CHF. Une étude de ces données montre aussi qu'un échantillon de 30 clients est suffisamment grand pour appliquer le TCL.

À la sortie du supermarché, on effectue un sondage auprès de 50 clients, en leur demandant le montant de leurs achats. On calcule ensuite la moyenne \bar{x} des 50 réponses collectées.

a) Le TCL s'applique-t-il ?

b) Quelle est la probabilité que la moyenne obtenue par ce sondage soit supérieure à 270 CHF ?

c) Quelle est la probabilité que la moyenne obtenue par ce sondage se situe à moins de 10 CHF de la vraie moyenne ?

d) Si on néglige les valeurs ayant moins de 0,3% de chances de se produire, entre quelles valeurs devrait se situer \bar{x} ?

5.3 Intervalles de confiance

Dans les exercices d'application du TCL de la section précédente, la moyenne μ de la variable statistique X pour la population est donnée. Si cela n'est pas le cas, il est tout de même possible d'estimer μ grâce au TCL et à la moyenne \bar{X} de variables statistiques X_1, X_2, \dots, X_n mesurant une caractéristique d'échantillons d'une même population.

Remarque

Le même problème se pose pour l'écart-type σ : si l'écart-type de la population n'est pas connu, quelle valeur utiliser pour le TCL ? En pratique, ce problème se révèle moins délicat que pour la moyenne, et nous nous permettrons d'utiliser l'écart-type de l'échantillon directement à la place de l'écart-type de la population⁷.

Première approche

Supposons que l'on cherche à estimer la taille moyenne μ des femmes suisses de 18 ans. Une enquête recueille les tailles x_1, x_2, \dots, x_{30} de 30 jeunes femmes suisses choisies au hasard ; la moyenne est $\bar{x} = 162,3$ cm.

Puisque cette moyenne \bar{x} est celle d'un échantillon, on ne peut pas affirmer que $\mu = 162,3$, c'est-à-dire que la taille moyenne de toutes les femmes suisses vaut 162,3 cm. Par contre, il est raisonnable d'affirmer que

la taille moyenne des femmes suisses de 18 ans vaut *environ* 162,3 cm.

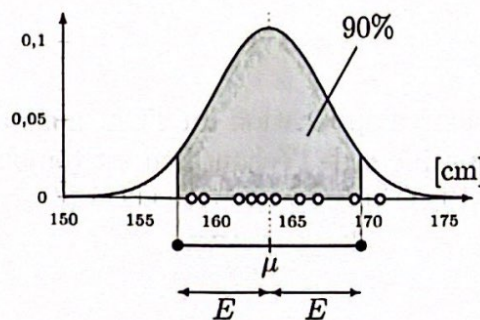
Une telle estimation s'appelle une **estimation ponctuelle**, c'est-à-dire une estimation par une seule valeur.

Approfondissement

Le problème avec une estimation ponctuelle est qu'on ne sait rien sur l'erreur que l'on risque de commettre. La vraie moyenne μ s'éloigne-t-elle de quelques millimètres, de quelques centimètres ou de quelques dizaines de centimètres de cette mesure ?

On va donc préférer estimer μ à l'aide d'un **intervalle de confiance**.

Ci-dessous, on a représenté en abscisses les moyennes \bar{x} de 10 enquêtes portant chacune sur la taille de 30 jeunes femmes suisses.



7. En théorie, on introduit parfois l'écart-type corrigé $\hat{\sigma}$ des mesures x_1, x_2, \dots, x_n , pour remplacer σ , où

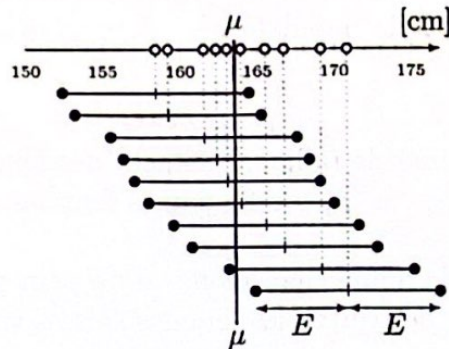
$$\hat{\sigma}^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

Pour se faire une idée de la situation, on suppose dans un premier temps que la moyenne μ ainsi que l'écart-type σ réels sont connus, et que les tailles suivent une loi normale. Selon le TCL, on a $\bar{X} \sim \mathcal{N}(\mu; \sigma^2/30)$, et E est calculé de telle sorte que

$$P(\mu - E \leq \bar{X} \leq \mu + E) = 90\%.$$

Comme le suggère la figure, environ 90% des moyennes \bar{x} vont appartenir à l'intervalle $[\mu - E; \mu + E]$.

En prenant les moyennes \bar{x} comme points de référence au lieu μ , on peut réinterpréter cette affirmation en disant que, **dans 90% des enquêtes**, la vraie moyenne μ est à une distance E ou moins de chaque \bar{x} :



Chacun des intervalles représenté autour d'une moyenne \bar{x} est un **intervalle de confiance** au niveau 90%. C'est-à-dire qu'en tirant 100 de ces intervalles au hasard, 90 en moyenne contiendront la vraie valeur de μ .

Le nombre E s'appelle la **marge d'erreur** de l'intervalle.

Terminologie

Plutôt que de dire

90 sur 100 des intervalles de confiance construits comme nous venons de le faire contiendront la moyenne μ

on dira⁸

μ appartient à l'intervalle de confiance avec un niveau de confiance de 90%

Remarque

Puisque nous nous à des situations d'application du TCL, nous allons admettre que, dans tous les exercices qui suivent, la taille n de l'échantillon est toujours "suffisamment grand".

8. Dans l'exemple des tailles des jeunes femmes suisses, μ est la vraie moyenne et ne varie donc pas ; de même, la moyenne \bar{x} et la marge d'erreur, une fois calculées à partir de l'échantillon, ne varient pas non plus. La moyenne μ appartient ou n'appartient pas à l'intervalle de confiance, il n'y a pas de probabilité en jeu. Il serait abusif de dire " μ appartient à l'intervalle de confiance avec une probabilité de 90%", d'où l'introduction d'une terminologie spécifique.

Exemple

Reprenons l'exemple des 30 femmes mesurant en moyenne $\bar{x} = 162,3$ cm (en pratique, on essaiera tout de même d'avoir un échantillon plus grand!), et supposons avoir un écart-type $\sigma = 12,3$ cm.

a) Dans ce problème que représentent les variables statistiques X et \bar{X} ?

b) Déterminez les paramètres de la loi normale \bar{X} .

c) Calculez l'intervalle de confiance correspondant au niveau 90%.