## *Estimation d'une moyenne à l'aide d'un échantillon*

### 1) Problématique de l'estimation :

Exemple :

On cherche à estimer la taille moyenne $m$ des hommes adultes dans une ville de 100 000 habitants. Pour cela, on sélectionne au hasard un échantillon de 1000 personnes de cette population et on évalue la taille moyenne $m_{ech}$ des individus de cet échantillon. On prend alors cette valeur comme estimation de la taille moyenne des 100 000 habitants.

Une question se pose alors :

Comment savoir si $m_{ech}$ est proche de $m$ d'une façon satisfaisante ?

Cette question va être résolue de deux façons, selon que l'on considère l'écart-type $\sigma$ des tailles des 100 000 habitants comme connu ou non. Pour cela, les valeurs obtenues pour chaque individu de l'échantillon vont être considérées comme étant des valeurs prises par des variables aléatoires.

Voyons cela plus en détail.

### 2) Modélisation mathématique du problème de l'échantillonnage

On se donne une **population d'individus statistiques $\Omega$ d'effectif $N$** ($N$ = 100 000 dans l'exemple précédent) pour laquelle on s'intéresse à un **caractère $x$** ($x$ est la taille dans l'exemple précédent). L'expérience aléatoire qui consiste à tirer un individu au hasard et s'intéresser à son caractère définit une **variable aléatoire $Y$ d'espérance $m$, de variance $V$ et d'écart-type $\sigma$**.

Considérons alors un **échantillon de taille $n$** (dans l'exemple $n = 1000$) tiré au hasard dans la population. Dans un tel échantillon, il ne peut y avoir deux individus identiques. On dit qu'il est obtenu sans remise. Toutefois, les lois mathématiques avec des variables aléatoires créées pour de tels échantillons ne sont pas simples à manipuler. On fait alors l'hypothèse que **l'échantillon a été obtenu dans un tirage avec remise**, ce qui signifie, qu'à chacune des $n$ étapes de sélection d'un individu, chacun des $N$ individus a la même chance d'être choisi. Cela peut être considéré dès que $N$ est suffisamment grand devant $n$.

Notons $x_1, x_2, \dots, x_n$ les caractères des $n$ individus de l'échantillon et considérons qu'ils sont les valeurs prises par **$n$ variables aléatoires $X_1, X_2, \dots, X_n$ indépendantes dans leur ensemble et de même distribution que la variable $Y$**.

La moyenne des caractères de l'échantillon est :

$$m_{ech} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Elle est la valeur prise par la variable aléatoire notée $\overline{X}$ et qualifiée de moyenne empirique :

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

La variance des caractères de l'échantillon est :

$$\boldsymbol{V_{ech} = \frac{(x_1 - m_{ech})^2 + (x_2 - m_{ech})^2 + \cdots + (x_n - m_{ech})^2}{n} = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n} - {m_{ech}}^2}$$

Elle est la valeur prise par la variable aléatoire notée $S^2$ et qualifiée de variance empirique :

$$\boldsymbol{S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n} = \frac{{X_1}^2 + {X_2}^2 + \cdots + {X_n}^2}{n} - \bar{X}^2}$$

Propriétés :

$$\boldsymbol{E(\bar{X}) = m}$$

$$\boldsymbol{E(S^2) = V - \frac{V}{n} = \frac{n-1}{n}\sigma^2}$$

Preuves :

$$E(\bar{X}) = \frac{E(X_1 + X_2 + \cdots + X_n)}{n} = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} = \frac{m + m + \cdots + m}{n} = \frac{n\,m}{n} = m$$

Les variables $X_1, X_2, \ldots, X_n$ étant indépendantes dans leur ensemble, elles le sont en particulier deux à deux et leurs covariances deux à deux sont nulles. Ainsi :

$$V(\bar{X}) = \frac{V(X_1 + X_2 + \cdots + X_n)}{n^2} = \frac{V(X_1) + V(X_2) + \cdots + V(X_n)}{n^2} = \frac{V + V + \cdots + V}{n^2} = \frac{n\,V}{n^2} = \frac{V}{n} = \frac{\sigma^2}{n}$$

D'autre part :

$$S^2 = \frac{\big((X_1 - m) + (m - \bar{X})\big)^2 + \big((X_2 - m) + (m - \bar{X})\big)^2 + \cdots + \big((X_n - m) + (m - \bar{X})\big)^2}{n}$$

$$\frac{1}{n}\sum_{i=1}^{n}\big((X_i - m) + (m - \bar{X})\big)^2 = \frac{1}{n}\sum_{i=1}^{n}((X_i - m)^2 + 2\,(X_i - m)\,(m - \bar{X}) + (m - \bar{X})^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 + \frac{2}{n}\,(m - \bar{X})\sum_{i=1}^{n}(X_i - m) + \frac{1}{n}\sum_{i=1}^{n}(m - \bar{X})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 + \frac{2}{n}\,(m - \bar{X})\left(\sum_{i=1}^{n}X_i - n\,m\right) + \frac{1}{n}\sum_{i=1}^{n}(\bar{X} - m)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 + \frac{2}{n}\,(m - \bar{X})\,(n\,\bar{X} - n\,m) + \frac{1}{n}n\,(\bar{X} - m)^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - 2\,(\bar{X} - m)^2 + (\bar{X} - m)^2$$

Finalement :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\bar{X} - m)^2$$

Ainsi :

$$E(S^2) = \frac{1}{n}\sum_{i=1}^{n}E(X_i - m)^2 - E((\bar{X} - m)^2) = \frac{1}{n}\sum_{i=1}^{n}V(X_i) - V(\bar{X}) = \frac{1}{n}\,n\,V - \frac{V}{n} = V - \frac{V}{n}$$

**Remarque** :

La variable aléatoire moyenne empirique $\bar{X}$, dont la valeur est la moyenne des caractères de l'échantillon a pour espérance $m$, c'est à la dire la grandeur qu'on cherche à estimer. Cela est une qualité intéressante car de nombreuses variables aléatoires se distribuent selon des lois normales et les valeurs qu'elles prennent sur un grand nombre de réalisations se regroupent le plus souvent autour de leur espérance, ce qui fait qu'on a plus de chance en ne considérant qu'une seule réalisation de $\bar{X}$ d'avoir une valeur proche de $m$. **On dit alors de la variable $\bar{X}$ d'échantillonnage qu'elle est un estimateur non biaisé de la moyenne.**

En revanche, la variable variance empirique $S^2$ n'a pas pour espérance la variance de la population mais une valeur inférieure d'une quantité appelé **biais**. Cependant, lorsque la taille de l'échantillon $n$ devient grande, ce biais tend vers zéro. **On dit alors de la variable $S^2$ d'échantillonnage qu'elle est un estimateur biaisé de la variance**.

Afin d'obtenir un estimateur non biaisé de la variance, on corrige légèrement $S^2$ en considérant :

$$\boldsymbol{S^{*2} = \frac{n}{n-1}S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}}$$

En effet :

$$\boldsymbol{E(S^{*2}) = \frac{n}{n-1}E(S^2) = \frac{n}{n-1}\,\frac{n-1}{n}\,V = V = \sigma^2}$$

Une estimation sans biais de la variance sera donc :

$$\boldsymbol{\hat{V} = \frac{(x_1 - m_{ech})^2 + (x_2 - m_{ech})^2 + \cdots + (x_n - m_{ech})^2}{n-1}}$$

et pour l'écart-type , on prendra naturellement :

$$\boldsymbol{\hat{\sigma} = \sqrt{\hat{V}}}$$

**3) Intervalle de confiance de l'estimation d'une moyenne, l'écart-type étant connu**

On se sert du théorème central limite qui énonce que si $\boldsymbol{X_1, X_2, \ldots, X_n}$ **sont $n$ variables aléatoires réelles indépendantes dans leur ensemble et de même distribution d'espérance $m$ et d'écart-type $\sigma$ alors la variable moyenne empirique $\bar{X}$ qui leur est associée tend vers une distribution normale quand $n$ tend vers l'infini.**

**En pratique, on considère cette approximation acceptable à partir de $n = 30$**.

Rappelons alors que pour une variable $X$ suivant une loi normale d'espérance $m$ et d'écart-type $\sigma$, nous avons trois intervalles de fluctuations qui sont d'intérêt pratique :

L'intervalle de fluctuation à 90 % caractérisé par :

$$P(X \in [m - 1{,}64\,\sigma, m + 1{,}64\,\sigma\,]) \approx 0{,}90$$

celui à 95 % :

$$P(X \in [m - 1{,}96\,\sigma, m + 1{,}96\,\sigma\,]) \approx 0{,}95$$

et enfin celui à 99 % :

$$P(X \in [m - 2{,}58\,\sigma, m + 2{,}58\,\sigma\,]) \approx 0{,}99$$

Limitons nous au cas de l'intervalle de fluctuation à 95 %. La variable moyenne empirique $\bar{X}$ dont la loi est assimilée à une distribution normale pour $n \geq 30$, produit donc dans 95 % des échantillons en très grand nombre qu'on réaliserait, une valeur $m_{ech}$ qui se situe dans l'intervalle de fluctuation :

$$\left[m - 1{,}96\,\frac{\sigma}{\sqrt{n}}, m + 1{,}96\,\frac{\sigma}{\sqrt{n}}\right]$$

On suppose donc et on le traduit par l'expression, « avec une confiance de 95 % » ou bien « avec une probabilité de 0,95 » que la moyenne $m_{ech}$ mesurée dans l'échantillon que l'on a réalisé, est dans cet intervalle. Ainsi :

$$m_{ech} \in \left[m - 1{,}96\,\frac{\sigma}{\sqrt{n}}, m + 1{,}96\,\frac{\sigma}{\sqrt{n}}\right]$$

Ce qui équivaut à écrire :

$$|m_{ech} - m| \leq \frac{\sigma}{\sqrt{n}}$$

Donc :

$$\boldsymbol{m \in \left[m_{ech} - 1{,}96\,\frac{\sigma}{\sqrt{n}}, m_{ech} + 1{,}96\,\frac{\sigma}{\sqrt{n}}\right]}$$

Nous obtenons ainsi une fourchette d'estimation de $m$ avec une confiance de 95 % ce qu'on appelle un **intervalle de confiance de $m$ à un niveau de confiance de 95 %.**

### 4) Intervalle de confiance de l'estimation d'une moyenne, l'écart-type étant inconnu

Une première méthode consiste à remplacer dans l'intervalle de confiance précédent $\sigma$ par son estimation $\hat{\sigma}$. L'intervalle de confiance modifié devient alors :

$$\left[m_{ech} - 1{,}96\,\frac{\hat{\sigma}}{\sqrt{n}}, m_{ech} + 1{,}96\,\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

Toutefois, cette formule ne donnant aucune indication sur la qualité de l'estimation de l'écart-type, il est préférable de procéder autrement en utilisant une autre variable statistique qui est :

$$T_{n-1} = \frac{\bar{X} - m}{S}\sqrt{n-1}$$

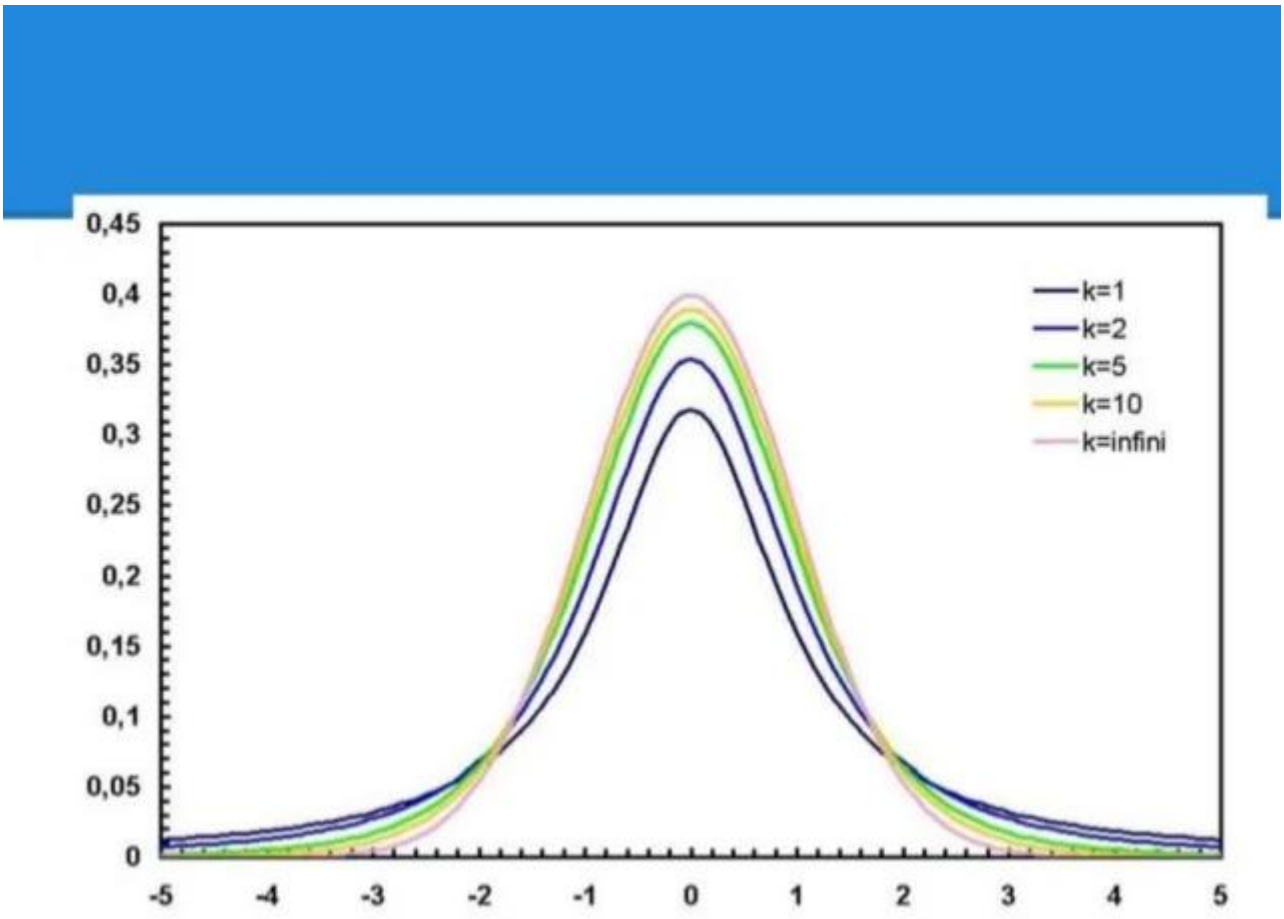
Où :

$$S = \sqrt{S^2}$$

On démontre que :

**$T_{n-1}$ est une variable aléatoire distribuée suivant une loi de Student à $n-1$ degrés de liberté**

L'allure de $T_k$ est la suivante pour différentes valeurs de $k$ est la suivante :




**Remarque** :

**Quand $k$ tend vers l'infini, la distribution de $T_k$ tend vers une loi normale centrée réduite**

On définit alors la valeur du réel $t > 0$ tel que :

$$P(T_{n-1} \in [-t, t]) = 0{,}95$$

Ce qui équivaut à trouver $t$ tel que :

$$P(T_{n-1} > t) = 0{,}025$$

On note pour cela cette valeur sous la forme $t_{0,025}$ qu'on appelle **quantile à 2,5 %, c'est-à-dire valeur qui a une probabilité de 2,5 % d'être dépassée par la variable $T_{n-1}$.**

L'intervalle de confiance à 95 % s'obtient en écrivant :

$$\frac{m_{ech} - m}{\sigma_{ech}} \sqrt{n-1} \in \left[-t_{0,025}, t_{0,025}\right]$$

Ce qui équivaut à :

$$\left|\frac{m_{ech} - m}{\sigma_{ech}} \sqrt{n-1}\right| \leq t_{0,025}$$

Soit :

$$|m_{ech} - m| \leq t_{0,025} \frac{\sigma_{ech}}{\sqrt{n-1}}$$

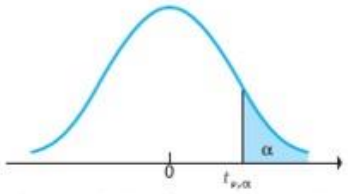
Soit finalement :

$$\boldsymbol{m \in \left[m_{ech} - t_{0,025} \frac{\sigma_{ech}}{\sqrt{n-1}}, m_{ech} + t_{0,025} \frac{\sigma_{ech}}{\sqrt{n-1}}\right]}$$

Ce qui définit l'intervalle de confiance à 95 %

Les valeurs de $t_{0,025}$ peuvent se lire dans une table comme celle qui suit

**Table 8** Upper Critical Values of Student's $t$ Distribution with $\nu$ Degrees of Freedom




For selected probabilities, $\alpha$, the table shows the values $t_{\nu,\alpha}$ such that $P(t_\nu > t_{\nu,\alpha}) = \alpha$, where $t_\nu$ is a Student's $t$ random variable with $\nu$ degress of freedom. For example, the probability is .10 that a Student's $t$ random variable with 10 degrees of freedom exceeds 1.372.

| | Probability of Exceeding the Critical Value | | | | | |
|---|---|---|---|---|---|---|
| $\nu$ | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.313 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.782 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.499 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.296 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.143 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.024 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.929 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 |
| 100 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 |
| $\infty$ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 |

*NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, September 2011.



Ainsi pour un échantillon de 30, le nombre de degrés de libertés du $T$ de Student est $n - 1 = 29$ et la valeur de $t_{0,025}$ vaut 2,045.

A noter que pour des valeurs de degrés de libertés supérieures ou égales à 100 les valeurs $t_\alpha$ données par la variable de Student sont proches de celles de la loi normale centrée réduite. Ainsi pour $t_{0,025}$ on se trouve proche de 1,96.