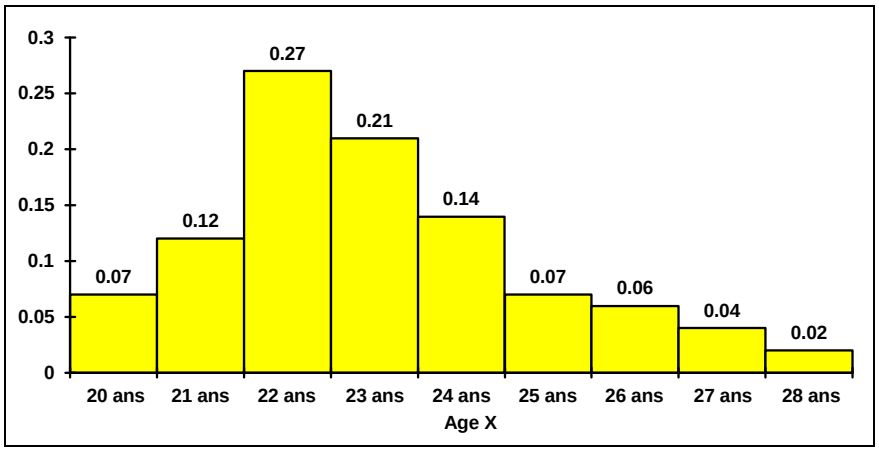
De la Statistique Descriptive à la Statistique Inférentielle, en passant par le Calcul des Probabilités.

1. Statistique Descriptive

On considère une population de taille ***N*** et la variable âge, notée ***X***, sur cette population.
La distribution de fréquences de ***X*** sur la population et sa représentation graphique sont données ci-après.

X	***f***
20	0.07
21	0.12
22	0.27
23	0.21
24	0.14
25	0.07
26	0.06
27	0.04
28	0.02
	100



On peut vérifier que la moyenne et la variance de ***X*** sont :

$$m = \sum f_k \, x_k = 23 \quad \text{et} \quad \sigma^2 = \sum f_k \left(x_k - 23\right)^2 = 3.48$$

2. Probabilités : échantillon de taille 1

On tire “au hasard” **un** individu de la population, c’est-à-dire que les ***N*** individus ont la même chance, $\frac{1}{N}$, d’être tirés (équiprobabilité).

Soit ***X*** l’âge de l’individu ; X_1 est une variable aléatoire réelle dont la distribution de probabilité est exactement la distribution de fréquences de ***X*** sur la population.
On a donc $\mathbf{E}(X_1) = 23$ et $\mathbf{V}(X_1) = 3.48$.

3. Probabilités : échantillon de taille 2

On tire “au hasard” deux individus de la population, successivement et avec remise, c’est-à-dire les N^2 couples d’individus de la population ont la même chance $\frac{1}{N^2}$ d’être tirés.

Soit X_1 et X_2 l’âge des deux individus et $\overline{X} = \frac{1}{2}(X_1 + X_2)$ l’âge moyen.

X_1 et X_2 sont deux variables aléatoires réelles, indépendantes et de même distribution de probabilité : la distribution de fréquences de X.

La loi de probabilité de $\overline{X}$ peut être calculée à partir du tableau suivant (les probabilités sont indiquées sous les valeurs de la variable.

X_1 \ X_2	20 0.07	21 0.12	22 0.27	23 0.21	24 0.14	25 0.07	26 0.06	27 0.04	28 0.02
20 0.07	20 $(0.07)^2$	20.5 0.07×0.12	21 0.07×0.27	21.5 0.07×0.21	**22** 0.07×0.14	22.5 0.07×0.07	23 0.07×0.06	23.5 0.07×0.04	24 0.07×0.02
21 0.12	20.5 0.12×0.07	21 $(0.12)^2$	21.5 0.12×0.27	**22** 0.12×0.21	22.5 0.12×0.14	23 0.12×0.07	23.5 0.12×0.06	24 0.12×0.04	24.5 0.12×0.02
22 0.27	21 0.27×0.07	21.5 0.27×0.12	**22** $(0.27)^2$	22.5 0.27×0.21	23 0.27×0.14	23.5 0.27×0.07	24 0.27×0.06	24.5 0.27×0.04	25 0.27×0.02
23 0.21	21.5 0.21×0.07	**22** 0.21×0.12	22.5 0.21×0.27	23 $(0.21)^2$	23.5 0.21×0.14	24 0.21×0.07	24.5 0.21×0.06	25 0.21×0.04	25.5 0.21×0.02
24 0.14	**22** 0.14×0.07	22.5 0.14×0.12	23 0.14×0.27	23.5 0.14×0.21	24 $(0.14)^2$	24.5 0.14×0.07	25 0.14×0.06	25.5 0.14×0.04	26 0.14×0.02
25 0.07	22.5 0.07×0.07	23 0.07×0.12	23.5 0.07×0.27	24 0.07×0.21	24.5 0.07×0.14	25 $(0.07)^2$	25.5 0.07×0.06	26 0.07×0.04	26.5 0.07×0.02
26 0.06	23 0.06×0.07	23.5 0.06×0.12	24 0.06×0.27	24.5 0.06×0.21	25 0.06×0.14	25.5 0.06×0.07	26 $(0.06)^2$	26.5 0.06×0.04	27 0.06×0.02
27 0.04	23.5 0.04×0.07	24 0.04×0.12	24.5 0.04×0.27	25 0.04×0.21	25.5 0.04×0.14	26 0.04×0.07	26.5 0.04×0.06	27 $(0.04)^2$	27.5 0.04×0.02
28 0.02	24 0.02×0.07	24.5 0.02×0.12	25 0.02×0.27	25.5 0.02×0.21	26 0.02×0.14	26.5 0.02×0.07	27 0.02×0.06	27.5 0.02×0.04	28 $(0.02)^2$

Les variables x_1 et x_2 étant indépendantes, on a, par exemple :

$\mathbf{P}(X_1 = 23 \text{ et } X_2 = 21) = \mathbf{P}(X_1 = 23) \times \mathbf{P}(X_2 = 21) = 0.21 \times 0.12$

et on a, dans ce cas, $\overline{X} = \frac{1}{2}(23 + 21) = 22$.

Il y a cinq manières distinctes d’obtenir $\overline{X} = 22$:

pour $X_1 = 24$ et $X_2 = 20$, pour $X_1 = 23$ et $X_2 = 21$, pour $X_1 = 22$ et $X_2 = 22$, pour $X_1 = 21$ et $X_2 = 23$, pour $X_1 = 20$ et $X_2 = 24$.

La probabilité d'avoir $\overline{X} = \mathbf{22}$ est donc égale à la somme de ces probabilités :

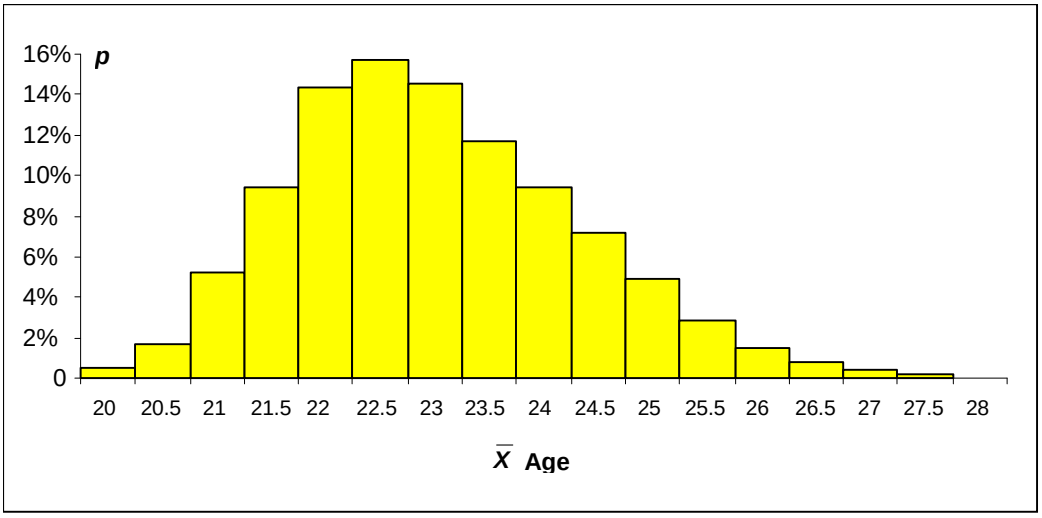
$$\mathbf{P}\left(\overline{X} = \mathbf{22}\right) = \mathbf{0.14} \times \mathbf{0.07} + \mathbf{0.21} \times \mathbf{0.12} + \mathbf{0.27}^{\mathbf{2}} + \mathbf{0.12} \times \mathbf{0.21} + \mathbf{0.07} \times \mathbf{0.14} = \mathbf{0.1429}$$

La distribution de probabilité de $\overline{X}$ et sa représentation graphique sont alors :

$\overline{X}$	20	20.5	21	21.5	22	22.5	23	23.5	24	24.5	25
p	0.0049	0.0168	0.0522	0.0942	0.1429	0.1568	0.1449	0.1166	0.0938	0.0712	0.0493

$\overline{X}$	25.5	26	26.5	27	27.5	28
p	0.028	0.0148	0.0076	0.004	0.0016	0.0004

Histogramme de fréquences de $\overline{X}$



On peut vérifier que l'espérance et la variance de $\overline{X}$ sont :

$$\mathbf{E}\left(\overline{X}\right) = \sum p_k \, \overline{x}_k = \mathbf{23} \quad \text{et} \quad \mathbf{V}\left(\overline{X}\right) = \sum p_k \left(\overline{x}_k - \mathbf{23}\right)^{\mathbf{2}} = \mathbf{174}$$

4. Probabilités : échantillon de taille 30

On tire “au hasard” un échantillon de 30 individus avec remise, c’est-à-dire les N^{30} échantillons possibles sont équiprobables.

Soit X_i l’âge du i-ème individu de l’échantillon et $\overline{X} = \frac{1}{30}\sum_{i=1}^{30} X_i$.

Alors $\overline{x}$ est une variable aléatoire réelle pour laquelle il est facile de montrer qu’elle a pour espérance 23 et pour variance 0.116 $\left(= 3.48/30\right)$.

On peut aussi montrer que sa distribution de probabilité est approximativement normale. C’est le théorème de la limite centrée. La variable aléatoire $Z = \frac{\overline{X} - 23}{\sqrt{0.116}}$ est alors une variable aléatoire, normale, centrée, réduite.
On lit dans la table de la distribution normale centrée réduite que la probabilité que Z soit compris entre -1.96 et 1.96 est égale à 0.95 :

$$\mathbf{P}(-1.96 < Z \leq 1.96) = 0.95$$

On en déduit :

$$\mathbf{P}\left(23 - 1.96\sqrt{0.116} < \overline{X} \leq 23 + 1.96\sqrt{0.116}\right) = 0.95$$

Ce résultat peut s’énoncer ainsi : 95% des N^{30} échantillons possibles de 30 individus ont un âge moyen compris entre $(23 - 0.67)$ ans et $(23 + 0.67)$ ans, c’est-à-dire entre 22.33 ans et 23.67 ans.

Ces résultats ne dépendent pas de la taille N de la population.
Ils sont valables que N soit égal à 100, à 100 000 ou à 50 millions !

5. Statistique inférentielle : estimation par intervalle de confiance

On suppose à présent que l'on ne connaît ni la distribution de fréquences, ni la moyenne, ni la variance de la variable âge sur la population.
De plus la taille de la population est trop grande pour que l'on puisse réaliser un "***recensement***", c'est-à-dire enquêter tous les individus de la population.

On décide alors d'effectuer un "***sondage***", c'est-à-dire d'enquêter une partie seulement de la population, appelée "***échantillon***" (de taille 30).

Estimation d'une moyenne.

L'objectif est d'estimer la moyenne m de X sur la population.
On sait que 95% des échantillons ont un âge moyen, $\bar{x}$, compris entre $m - 1{,}96\frac{\sigma}{\sqrt{30}}$ et $m + 1{,}96\frac{\sigma}{\sqrt{30}}$.
On peut donc dire aussi que m est compris entre $\bar{x} - 1{,}96\frac{\sigma}{\sqrt{30}}$ et $\bar{x} + 1{,}96\frac{\sigma}{\sqrt{30}}$ pour 95% des échantillons.

On tire un échantillon aléatoire de taille 30 et on calcule la moyenne $\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i$ et la variance "corrigée" $s^2 = \frac{1}{29}\sum_{i=1}^{30}(x_i - \bar{x})^2$ des âges sur l'échantillon.

De la même manière que l'on a montré que $\bar{x}$ est l'observation d'une variable aléatoire $\bar{X}$, de moyenne m et de variance σ^2/n, on peut montrer que s^2 est l'observation d'une variable aléatoire S^2 dont l'espérance est σ^2 $\left(E(S^2) = \sigma^2\right)$.

On propose alors l'estimation de m par intervalle de confiance à 95% :

$$\left[\bar{x} - 1{,}96\frac{s}{\sqrt{30}} \quad ; \quad \bar{x} + 1{,}96\frac{s}{\sqrt{30}}\right]$$

Supposons que l'on ait trouvé $\bar{x} = 22{,}52$ et $s^2 = 3{,}60 = (1{,}90)^2$.

L'estimation de m par intervalle de confiance à 95% est :

$$\left[22{,}52 - 1{,}96\frac{1{,}90}{\sqrt{30}} \quad ; \quad 22{,}52 + 1{,}96\frac{1{,}90}{\sqrt{30}}\right]$$

$$\left[21{,}84 \quad ; \quad 23{,}20\right]$$

Cet intervalle aléatoire "recouvre" bien la valeur fixe m (=23) supposée inconnue dans cette partie.

Estimation d'une proportion.

Dans le cas où l'on veut estimer la proportion p d'une partie A de la population, on peut vérifier qu'une « proportion » est la « moyenne » d'une variable réelle particulière Y, appelée ***indicatrice***.

L'indicatrice de A est la variable réelle définie sur la population qui prend la valeur $y_i = 1$ si l'individu i appartient à A et la valeur $y_i = 0$ sinon.

Alors la moyenne de Y sur la population $\frac{1}{N}\sum_{i=1}^{N} y_i$ est tout simplement la proportion p et la variance $\left(= \frac{1}{N}\sum_{i=1}^{N} y_i^2 - p^2\right)$ est égale à $p - p^2$ car $y_i^2 = y_i$ pour tout i.

> Estimer une proportion p revient donc à estimer la moyenne d'une variable réelle Y qui a pour écart-type $\sqrt{p(1-p)}$.

Si on note f la proportion de A dans l'échantillon, de même que l'on a $m = p$ et $\sigma^2 = m(1-p)$, on a : $\bar{x} = f$ et $s^2 = f(1-f)$

En résumé

Si on observe sur un échantillon de taille ***n*** une moyenne $\bar{x}$ et un écart-type ***s***, l'estimation de ***m*** par intervalle à 95% de confiance est :

$$\left[\bar{x} - 1{,}96 \frac{s}{\sqrt{n}} \; ; \; \bar{x} + 1{,}96 \frac{s}{\sqrt{n}} \right]$$

L'estimation à 95% de confiance de ***p*** est :

$$\left[f - 1{,}96 \sqrt{\frac{f(1-f)}{n}} \; ; \; f + 1{,}96 \sqrt{\frac{f(1-f)}{n}} \right]$$

Pour un intervalle de confiance à 90%, on remplace 1.96 par 1.645.

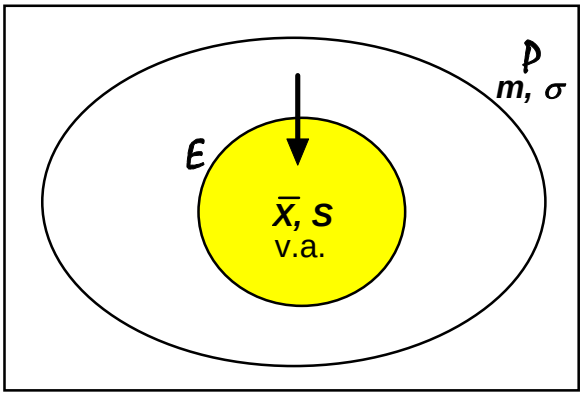
La statistique inférentielle procède donc en deux étapes.

1- dans la première, appelée ***échantillonnage***, on étudie les propriétés des valeurs observées sur ***tous les échantillons*** possibles de taille ***n*** en fonction des paramètres de la population,

2- dans la deuxième, appelée ***estimation***, on utilise les résultats précédents pour inférer à la population les valeurs observées sur ***l'échantillon*** qui a été tiré au hasard.

Echantillonnage

(de la population vers **les échantillons**)



Estimation

(**d'un échantillon** vers la population)