Chapitre 3

Distributions d'échantillonage

3.1 Généralités sur la notion d'échantillonnage

3.1.1 Population et échantillon

On appelle population la totalité des unités de n'importe quel genre prises en considération par le statisticien. Elle peut être finie ou infinie.

Un échantillon est un sous-ensemble de la population étudiée.

Qu'il traite un échantillon ou une population, le statisticien décrit habituellement ces ensembles à l'aide de mesures telles que le nombre d'unités, la moyenne, l'écart-type et le pourcentage.

- Les mesures que l'on utilise pour décrire une population sont des paramètres. Un paramètre est une caractéristique de la population.
- Les mesures que l'on utilise pour décrire un échantillon sont appelées des statistiques. Une statistique est une caractéristique de l'échantillon.

Nous allons voir dans ce chapitre et dans le suivant comment les résultats obtenus sur un échantillon peuvent être utilisés pour décrire la population. On verra en particulier que les statistiques sont utilisées pour estimer les paramètres.

Afin de ne pas confondre les statistiques et les paramètres, on utilise des notations différentes, comme le présente le tableau récapitulatif suivant.

	POPULATION	ÉCHANTILLON
DÉFINITION	C'est l'ensemble des unités considérées par le statisticien.	C'est un sous-ensemble de la population choisie pour étude.
CARACTÉRISTIQUES	Ce sont les paramètres	Ce sont les statistiques
NOTATIONS	N = taille de la population (si elle est finie)	n = taille de l'échantillon
Si on étudie un caractère quantitatif	moyenne de la population $m = \frac{1}{N}\sum_{i=1}^{N} x_i$	moyenne de l'échantillon $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
	écart-type de la population $\sigma_{pop} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2}$	écart-type de l'échantillon $\sigma_{ech} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Si on étudie un qualitatif	proportion dans la population p	proportion dans l'échantillon f

3.1.2 L'échantillonage

Avantages de l'échantillonnage

- Coût moindre.
- Gain de temps.
- C'est la seule méthode qui donne des résultats dans le cas d'un test destructif.

Méthodes d'échantillonnage

- Échantillonnage sur la base du jugement (par exemple, dans les campagnes électorales certains districts électoraux sont des indicateurs fiables de l'opinion publique).
- Échantillonnage aléatoire simple. Tous les échantillons possibles de même taille ont la même probabilité d'être choisis et tous les éléments de la population ont une chance égale de faire partie de l'échantillon (On utilise souvent une table de nombres aléatoires pour s'assurer que le choix des éléments s'effectue vraiment au hasard).

Remarque 1 *Il existe deux autres méthodes d'échantillonnage aléatoire mais elles ne nous intéressent pas ici . Ce sont l'échantillonnage stratifié et l'échantillonnage par grappes.*

Bien entendu, seul l'échantillonnage aléatoire nous permettra de juger objectivement de la valeur des estimations faites sur les caractéristiques de la population.

Inconvénient de l'échantillonnage

L'échantillonnage a pour but de fournir suffisamment d'informations pour pouvoir faire des déductions sur les caractéristiques de la population. Mais bien entendu, les résultats obtenus d'un échantillon à l'autre vont être en général différents et différents également de la valeur de la caractéristique correspondante dans la population. On dit qu'il y a des fluctuations d'échantillonnage. Comment, dans ce cas, peut-on tirer des conclusions valables ? En déterminant les lois de probabilités qui régissent ces fluctuations. C'est l'objet de ce chapitre.

3.2 La variable aléatoire : moyenne d'échantillon

3.2.1 Introduction

Position du problème :

Si nous prélevons un échantillon de taille n dans une population donnée, la moyenne de l'échantillon nous donnera une idée approximative de la moyenne de la population. Seulement si nous prélevons un autre échantillon de même taille, nous obtiendrons une autre moyenne d'échantillon. Sur l'ensemble des échantillons possibles, on constatera que certains ont une moyenne proche de la moyenne de la population et que d'autres ont une moyenne qui s'en écarte davantage.

Comment traiter le problème ?

Un échantillon de taille n (appelé aussi un n-échantillon), obtenu par échantillonnage aléatoire, va être considéré comme le résultat d'une expérience aléatoire. A chaque échantillon de taille n on peut associer la valeur moyenne des éléments de l'échantillon. On a donc défini une variable aléatoire qui à chaque n-échantillon associe sa moyenne échantillonnale. On la note $\bar{X}$. Cette variable aléatoire possède bien entendu :

- Une distribution de probabilité.
- Une valeur moyenne (la moyenne des moyennes d'échantillons, vous suivez toujours ?).
- Un écart-type.

Le but de ce paragraphe est de déterminer ces trois éléments.

Avant de continuer, essayons de comprendre sur un exemple ce qui se passe.

Exemple 2 *Une population est constituée de 5 étudiants en statistique (le faible effectif n'est pas dû à un manque d'intérêt pour la matière de la part des étudiants mais au désir de ne pas multiplier inutilement les calculs qui vont suivre !). Leur professeur s'intéresse au temps hebdomadaire consacré à l'étude des statistiques par chaque étudiant.*

On a obtenu les résultats suivants.

Étudiant	Temps d'étude (en heures)
A	7
B	3
C	6
D	10
E	4
Total	30

La moyenne de la population est $m = 30/5 = 6$.

Si le professeur choisit un échantillon de taille 3, quelles sont les différentes valeurs possibles pour la moyenne de son échantillon ? Quelle relation existe-t-il entre cette moyenne d'échantillon et la véritable moyenne 6 de la population ?

Toutes les possibilités sont regroupées dans le tableau ci-dessous.

Numéro de l'échantillon	Échantillon	Valeurs du temps d'étude dans cet échantillon	Moyennes de l' échantillon
1	A, B, C	7,3,6	5.33
2	A, B, D	7,3,10	6.67
3	A, B, E	7,3,4	4.67
4	A, C, D	7,6,10	7.67
5	A, C, E	7,6,4	5.67
6	A, D, E	7,10,4	7.00
7	B, C, D	3,6,10	6.33
8	B, C, E	3,6,4	4.33
9	B, D, E	3,10,4	5.67
10	C, D, E	6,10,4	6.67
Total			60.00

On constate que :
- Il y a 10 échantillons ($C_5^3 = 10$).
- La moyenne des échantillons varie entre 4.33 et 7.67, ce qui signifie que la distribution des moyennes d'échantillon est moins dispersée que la distribution des temps d'étude des étudiants, située entre 3 et 10.
- Il est possible que deux échantillons aient la même moyenne. Dans cet exemple, aucun n'a la moyenne de la population ($m = 6$).
- La moyenne des moyennes d'échantillon est $E(\bar{X}) = 60/10 = 6$.

En fait, nous allons voir que le fait que l'espérance de $\bar{X}$ (c'est-à-dire la moyenne des moyennes d'échantillon) est égale à la moyenne de la population n'est pas vérifié seulement dans notre exemple. C'est une loi générale.

Bien, me direz-vous, mais pourquoi faire tout cela ? Dans la réalité, on ne choisit qu'un seul échantillon. Alors comment le professeur de statistique qui ne connaît qu'une seule moyenne d'échantillon pourra-t-il déduire quelque chose sur la moyenne de la population ? Tout simplement en examinant "jusqu'à quel point" la moyenne d'un échantillon unique s'approche de la moyenne de la population. Pour cela, il lui faut la distribution théorique de la variable aléatoire $\bar{X}$ ainsi que l'écart-type de cette distribution.

3.2.2 Etude de la variable : moyenne d'échantillon

Définition de la variable

On considère une population dont les éléments possèdent un caractère mesurable qui est la réalisation d'une variable aléatoire X qui suit une loi de probabilité d'espérance m et d'écart-type σ_{pop}. On suppose que la population est infinie ou si elle est finie que l'échantillonnage se fait avec remise.
- On prélève un échantillon aléatoire de taille n et on mesure les valeurs de X sur chaque élément de l'échantillon. On obtient une suite de valeurs $x_1, x_2, \ldots, x_n$.
- Si on prélève un deuxième échantillon toujours de taille n, la suite des valeurs obtenues est $x'_1, x'_2, \ldots, x'_n$, puis $x''_1, x''_2, \ldots, x''_n$... etc... pour des échantillons supplémentaires.

x_1, x'_1, x''_1,... peuvent être considérées comme les valeurs d'une variable aléatoire X_1 qui suit la loi de X. De même, x_2, x'_2, x''_2,... peuvent être considérées comme les valeurs d'une variable aléatoire X_2 qui suit aussi la loi de X, ... et

x_n, x'_n, x''_n,... celles d'une variable aléatoire X_n qui suit encore et toujours la même loi, celle de X.

- X_1 pourrait se nommer "valeur du premier élément d'un échantillon". X_2 pourrait se nommer "valeur du deuxième élément d'un échantillon". ....X_n pourrait se nommer "valeur du n-ième élément d'un échantillon".
- L'hypothèse d'une population infinie ou d'un échantillonnage avec remise nous permet d'affirmer que ces n variables aléatoires sont indépendantes.

Rappel sur les notations : Par convention, on note toujours les variables aléatoires à l'aide de lettres majuscules (X_i) et les valeurs qu'elles prennent dans une réalisation à l'aide de lettres minuscules (x_i).

Si les valeurs prises par X dans un échantillon sont $x_1, x_2, \ldots, x_n$, la moyenne $\bar{x}$ de l'échantillon est donnée par $\bar{x} = \dfrac{1}{n}(x_1 + \cdots + x_n) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$. Cette valeur n'est rien d'autre que la valeur prise dans cet échantillon de la variable aléatoire $\dfrac{1}{n}(X_1 + \cdots + X_n) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.

Définition 3 *On définit donc la* variable aléatoire *moyenne d'échantillon* $\bar{X}$ *par*

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Paramètres descriptifs de la distribution

On applique les propriétés de l'espérance et de la variance étudiées au chapitre 2.

- $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{nm}{n} = m$, car les variables suivent toutes la même loi d'espérance m.
- $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2_{pop}}{n^2} = \frac{\sigma^2_{pop}}{n}$, car les variables suivent toutes la même loi de variance et sont indépendantes.

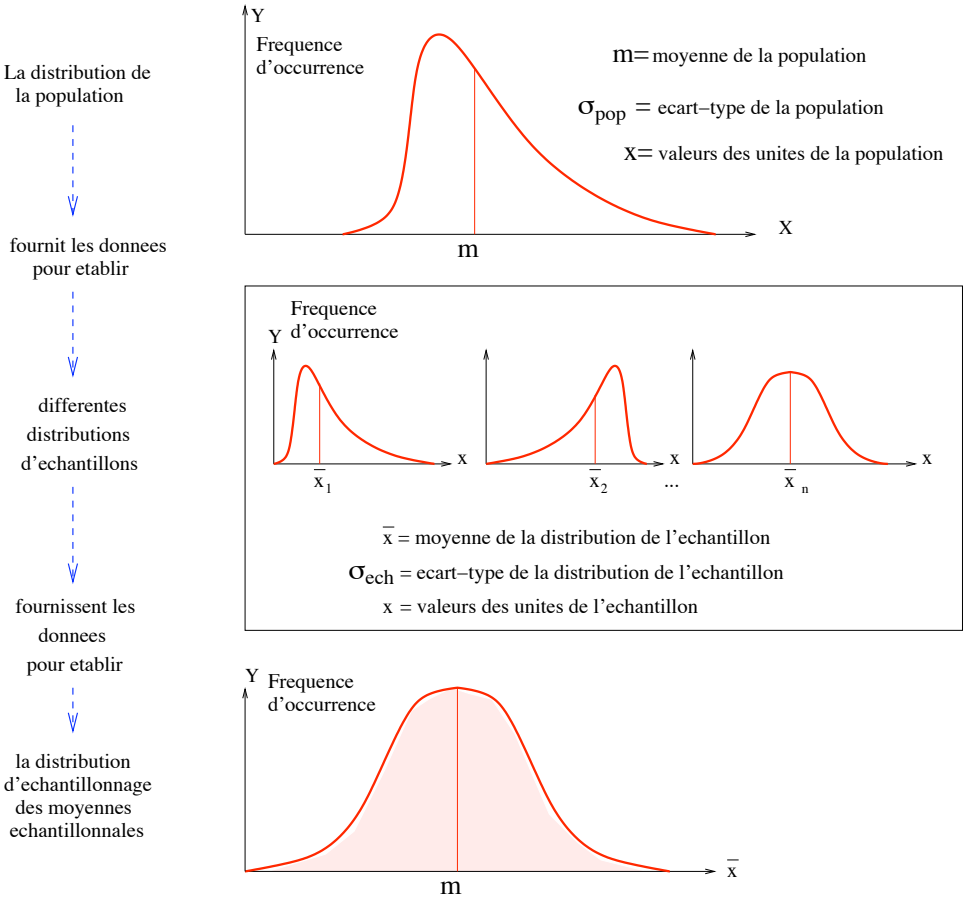
Proposition 4

$$E(\bar{X}) = m, \quad Var(\bar{X}) = \frac{\sigma^2_{pop}}{n}.$$

Remarque 5
1. *Nous venons de démontrer ce que nous avions constaté sur notre exemple : la moyenne de la distribution d'échantillonnage des moyennes est égale à la moyenne de la population.*
2. *On constate que plus n croît, plus* $Var(\bar{X})$ *décroît.*

Dans l'exemple d'introduction, nous avions en effet constaté que la distribution des moyennes d'échantillon était moins dispersée que la distribution initiale. En effet, à mesure que la taille de l'échantillon augmente, nous avons accès à une plus grande quantité d'informations pour estimer la moyenne de la population. Par conséquent, la différence probable entre la vraie valeur de la moyenne de la population et la moyenne échantillonnale diminue. L'étendue des valeurs possibles de la moyenne échantillonnale diminue et le degré de dispersion de la distribution aussi. $\sigma(\bar{X})$ est aussi appelé l'*erreur-type* de la moyenne.

On peut schématiser le passage de la distribution de la variable aléatoire X à celle de la variable aléatoire $\bar{X}$ en passant par les différents échantillons par le graphique ci-après.



m= moyenne de la distribution d'echantillonnage des moyennes

$\sigma_{pop/\sqrt{n}}$ = ecart–type de la distribution d'echantillonnage des moyennes

$\bar{x}$ = valeurs des moyennes de tous les echantillons possibles

Mais connaître les paramètres descriptifs de la distribution de $\bar{X}$ ne suffit pas. Il faut connaître aussi sa distribution de probabilité. On se demande alors : dépend elle

1. de la distribution de X ?
2. de la taille n de l'échantillon ?

3.3 La variable aléatoire : variance d'échantillon

La variance $\sigma^2_{ech} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ d'un n-échantillon est la réalisation de la variable aléatoire $\Sigma^2_{ech} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. On peut se demander si cette variable possède la même propriété que la variable moyenne d'échantillon, c'est-à-dire si l'espérance de Σ^2_{ech} est égale à la variance de la population.

3.3.1 Espérance de la variable aléatoire Σ^2_{ech}

Autre expression de Σ^2_{ech}

$$\begin{aligned}
\Sigma^2_{ech} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - m + m - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 + \frac{2}{n}\sum_{i=1}^{n}(X_i - m)(m - \bar{X}) + \frac{1}{n}\sum_{i=1}^{n}(m - \bar{X})^2 \\
&= A + B + C,
\end{aligned}$$

Or, $B = \frac{2}{n}(m - \bar{X})\sum_{i=1}^{n}(X_i - m) = -2(m - \bar{X})^2$ et $C = (m - \bar{X})^2$. On en déduit que

$$\Sigma^2_{ech} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\bar{X} - m)^2.$$

Espérance de Σ^2_{ech}

Proposition 6

$$E(\Sigma^2_{ech}) = \frac{n-1}{n}\sigma^2_{pop}.$$

Preuve.

$$\begin{aligned}
E(\Sigma^2_{ech}) &= \frac{1}{n}\sum_{i=1}^{n}E((X_i - m)^2) - E((\bar{X} - m)^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}Var(X_i) - Var(\bar{X}) \\
&= \sigma^2_{pop} - \frac{1}{n}\sigma^2_{pop} = \frac{n-1}{n}\sigma^2_{pop}.
\end{aligned}$$

Conclusion. La moyenne des variances d'échantillon n'est pas la variance de la population, mais la variance de la population multipliée par $\frac{n-1}{n}$. Bien sûr, si n est très grand, ces deux nombres seront très proches l'un de l'autre.

3.3.2 La variable aléatoire S^2

Définition de S^2

Pour pouvoir déterminer une valeur approchée de σ^2_{pop} et savoir quelle erreur on commet en effectuant cette approximation, on veut disposer d'une variable dont l'espérance est la variance de la population. Nous allons donc considérer une nouvelle variable aléatoire S^2.

Définition 7

$$S^2 = \frac{n}{n-1}\Sigma^2_{ech} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

On a bien entendu $E(S^2) = \sigma^2_{pop}$.

Nous verrons plus tard que cela signifie que S^2 est un estimateur *sans biais* de σ^2_{pop}.

Distribution de S^2

Nous supposons ici que X suit une loi normale.

On considère la variable $Y = \dfrac{n\Sigma^2_{ech}}{\sigma^2_{pop}} = \displaystyle\sum_{i=1}^{n}(\frac{X_i - \bar{X}}{\sigma_{pop}})^2$.

Y est une somme d'écarts réduits relatifs à une variable normale. D'après ce que nous avons vu au chapitre 3, paragraphe 7.1, nous pouvons affirmer que Y suit une loi du χ^2 à $n-1$ degrés de liberté (on perd un degré de liberté car on a estimé le paramètre m par $\bar{X}$).

Proposition 8 *$Y = \dfrac{(n-1)S^2}{\sigma^2_{pop}}$ suit une loi χ^2_{n-1}.*

Remarque 9 *Encore une fois, on n'a pas directement la loi de S^2 mais celle de $\frac{(n-1)S^2}{\sigma^2_{pop}}$.*

Approximation de la distribution de S^2 dans le cas des grands échantillons : $n \geq 30$

Nous avons vu au chapitre 3 que lorsque n est grand ($n \geq 30$), on pouvait approcher la loi χ^2_ν par la loi $\mathcal{N}(\nu, \sqrt{2\nu})$. Donc Y suit approximativement une loi normale, $E(Y) \simeq n-1$ et $Var(Y) \simeq 2(n-1)$.

Proposition 10 *Si $n \geq 30$, S^2 suit une loi $\mathcal{N}(\sigma^2_{pop}, \sigma^2_{pop}\sqrt{\dfrac{2}{n-1}})$, en première approximation.*

Preuve. La loi de S^2 est alors approximativement normale, son espérance vaut σ^2_{pop} et sa variance approximativement

$$Var(S^2) = Var(\frac{\sigma^2_{pop}}{n-1}Y) = \frac{\sigma^4_{pop}}{(n-1)^2}Var(Y) \simeq \frac{2\sigma^4_{pop}}{n-1}.$$

3.4 Distribution de la moyenne d'échantillon

Nous allons distinguer deux cas : celui des grands échantillons ($n \geq 30$) et celui des petits échantillons ($n < 30$).

3.4.1 Cas des grands échantillons : $n \geq 30$.

On peut appliquer le théorème centrale-limite.

1. Nous sommes en présence de n variables aléatoires indépendantes.
2. Elles suivent la même loi d'espérance m et de variance σ^2_{pop}, donc aucune n'est prépondérante.

Conclusion. Lorsque n devient très grand, la distribution de $S = X_1 + \cdots + X_n$ se rapproche de celle de la loi normale d'espérance nm et de variance $n\sigma^2_{pop}$, S suit approximativement $\mathcal{N}(nm, n\sigma^2_{pop})$.

Par conséquent, pour n assez grand, la distribution de $\bar{X} = S/n$ se rapproche de celle de la loi normale d'espérance m et de variance σ^2_{pop}/n c'est-à-dire $\mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}})$. On peut donc considérer que $\dfrac{\bar{X} - m}{\sigma_{pop}/\sqrt{n}}$ suit la loi $\mathcal{N}(0, 1)$.

Proposition 11 *Si $n \geq 30$, $\bar{X}$ suit approximativement $\mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}})$.*

Remarque 12
- *En pratique, on considère que cela est vrai à partir de $n \geq 30$ et que lorsque la forme de la distribution de X est pratiquement symétrique, $n \geq 15$ est convenable.*
- *Ce théorème est très puissant car il n'impose aucune restriction sur la distribution de X dans la population.*
- *Si la variance est inconnue, un grand échantillon ($n \geq 30$) permet de déduire une valeur fiable pour σ^2_{pop} en calculant la variance de l'échantillon σ^2_{ech} et en posant*

$$\sigma^2_{pop} = \frac{n}{n-1}\sigma^2_{ech} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

comme on l'a vu au paragraphe précédent.

3.4.2 Cas des petits échantillons : $n < 30$

Nous nous plaçons alors exclusivement dans le cas où X suit une loi normale dans la population.

Nous allons encore distinguer deux cas : celui où σ_{pop} est connu et celui où σ_{pop} est inconnu.

Cas où σ_{pop} est connu

X suit une loi normale $\mathcal{N}(m, \sigma_{pop})$ donc les variables X_i suivent toutes la même loi $\mathcal{N}(m, \sigma_{pop})$. De plus elles sont indépendantes. D'après la propriété vue au chapitre 3 sur la somme de lois normales indépendantes, $S = X_1 + \cdots + X_n$ a une distribution normale et la variable $\bar{X} = S/n$ suit aussi une loi normale, la loi $\mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}})$. Donc $\dfrac{\bar{X} - m}{\sigma_{pop}/\sqrt{n}}$ suit la loi $\mathcal{N}(0, 1)$.

$$\boxed{\text{Si } \begin{cases} n < 30 \\ \sigma_{pop} \text{ connu} \end{cases}, \bar{X} \text{ suit } \mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}}).}$$

Cas où σ_{pop} est inconnu

Lorsque l'échantillonnage s'effectue à partir d'une population normale de variance inconnue et que la taille de l'échantillon est petite ($n < 30$), l'estimation de la variance effectuée au paragraphe précédent n'est plus fiable. On ne peut plus écrire $\sigma^2_{pop} \simeq \frac{n}{n-1}\sigma^2_{ech}$ car σ^2_{ech} varie trop d'échantillon en échantillon.

L'écart-type de la distribution de $\bar{X}$ n'est donc plus une constante $\frac{\sigma_{pop}}{\sqrt{n}}$ connue approximativement grâce à $\frac{\sigma_{pop}}{\sqrt{n}} \simeq \frac{\sigma_{ech}}{\sqrt{n-1}}$. On va alors considérer que l'écart-type de $\bar{X}$ sera donné dans chaque échantillon par une valeur différente de $\frac{\sigma_{ech}}{\sqrt{n-1}}$.

Nous devons donc considérer σ_{ech} comme la réalisation d'une variable aléatoire, la variable écart-type d'échantillon, notée Σ_{ech} et définie par $\Sigma^2_{ech} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

La variable aléatoire $T = \dfrac{\bar{X} - m}{\Sigma_{ech}/\sqrt{n-1}} = \dfrac{\sqrt{n-1}(\bar{X} - m)}{\Sigma_{ech}}$ ne suit plus alors une loi normale car le dénominateur n'est pas une constante.

En divisant numérateur et dénominateur par σ_{pop}, on écrit T sous la forme

$$T = \frac{\sqrt{n-1}(\bar{X} - m)}{\Sigma_{ech}} = \frac{\sqrt{n-1}\frac{\bar{X}-m}{\sigma_{pop}/\sqrt{n}}}{\sqrt{\sum_{i=1}^{n}(\frac{X_i - \bar{X}}{\sigma_{pop}})^2}}.$$

On reconnaît au numérateur une variable aléatoire qui suit une loi $\mathcal{N}(0,1)$, multipliée par un facteur $\sqrt{n-1}$, et au dénominateur une somme de carrés de variables suivant aussi la loi $\mathcal{N}(0,1)$. Le carré du dénominateur suit donc une loi du χ^2. Mais quel est son nombre de degrés de liberté ?

Le concept de degré de liberté

- Pourquoi chercher le nombre de degrés de liberté ?
  Pour pouvoir utiliser correctement les tables de lois de probabilité qui dépendent d'un nombre de degrés de liberté (en particulier pour la distribution de Student et celle du χ^2).
- Que représente le nombre de degrés de liberté ?
  C'est une quantité qui est toujours associée à une somme de carrés et qui représente le nombre de carrés indépendants dans cette somme.
- Comment calculer le nombre de degrés de liberté ?
  Il y a deux façons de procéder.
  1. Soit on effectue la différence entre le nombre total de carrés et le nombre de relations qui lient les différents éléments de la somme.
  2. Soit on effectue la différence entre le nombre total de carrés et le nombre de paramètres que l'on doit estimer pour effectuer le calcul.

Dans le cas de notre somme $\sum_{i=1}^{n}(\frac{X_i - \bar{X}}{\sigma_{pop}})^2$, envisageons les deux façons de compter le nombre de degrés de liberté.

1. Le nombre de carrés dans la somme est n. Il y a une relation entre les variables $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$. Le nombre de degrés de liberté est donc $n-1$.
2. Le nombre de carrés dans la somme est n. Lorsqu'on dit que $\sum_{i=1}^{n}(\frac{X_i-\bar{X}}{\sigma_{pop}})^2$ est une somme de carrés de variables normales centrées réduites, on remplace m par $\bar{X}$. On a donc estimé un paramètre. On trouve encore que le nombre de degrés de liberté est $n-1$.

Proposition 13 *Si* $\begin{cases} n < 30 \\ \sigma_{pop} \text{ connu} \end{cases}$ *, la variable* $T = \dfrac{\bar{X} - m}{\Sigma_{ech}/\sqrt{n-1}}$ *suit une loi de Student à* $n-1$ *degrés de liberté, notée* T_{n-1}.

Revoir éventuellement la définition de la loi de Student dans le chapitre précédent.

Remarque 14 *Dans ce dernier cas (petits échantillons et X suit une loi normale de variance inconnue), on ne trouve pas directement la loi suivie par mais celle suivie par* $T = \dfrac{\bar{X} - m}{\Sigma_{ech}/\sqrt{n-1}}$.

Exercice 15 *Le responsable d'une entreprise a accumulé depuis des années les résultats à un test d'aptitude à effectuer un certain travail. Il semble plausible de supposer que les résultats au test d'aptitude sont distribués suivant une loi normale de moyenne* $m = 150$ *et de variance* $\sigma^2_{pop} = 100$. *On fait passer le test à 25 individus de l'entreprise. Quelle est la probabilité que la moyenne de l'échantillon soit entre 146 et 154 ?*

Solution de l'exercice 15. *Test d'aptitude.*

On considère la variable aléatoire $\bar{X}$ moyenne d'échantillon pour les échantillons de taille $n = 25$. On cherche à déterminer $P(146 < \bar{X} < 154)$.

Pour cela, il nous faut connaître la loi suivie par $\bar{X}$. Examinons la situation. Nous sommes en présence d'un petit échantillon ($n < 30$) et heureusement dans le cas où la variable X (résultat au test d'aptitude) suit une loi normale. De plus, σ_{pop} est connu. Donc $\bar{X}$ suit $\mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}}) = \mathcal{N}(150, 10/5)$. On en déduit que $T = \frac{\bar{X}-150}{2}$ suit $\mathcal{N}(0,1)$.

La table donne

$$\begin{aligned} P(146 < \bar{X} < 154) &= P(\frac{146-150}{2} < T < \frac{154-150}{2}) = P(-2 < T < 2) \\ &= 2P(0 < T < 2) = 2 \times 0.4772 = 0.9544. \end{aligned}$$

3.5 Distribution de la variable proportion d'échantillon

Il arrive fréquemment que nous ayions à estimer dans une population une proportion p d'individus possédant un caractère qualitatif donné.

Bien sûr, cette proportion p sera estimée à l'aide des résultats obtenus sur un n-échantillon. La proportion f obtenue dans un n-échantillon est la valeur observée d'une variable aléatoire F, fréquence d'apparition de ce caractère dans un échantillon de taille n, appelée proportion d'échantillon. On se pose une troisième fois la question. La moyenne des fréquences d'observation du caractère sur l'ensemble de tous les échantillons de taille n est-elle égale à la proportion p de la population ?

3.5.1 Paramètres descriptifs de la distribution de F

F est la fréquence d'apparition du caractère dans un échantillon de taille n. Donc $F = X/n$ où X est le nombre de fois où le caractère apparaît dans le n-échantillon.

Par définition X suit $\mathcal{B}(n, p)$. Donc $E(F) = np$ et $Var(F) = npq$.

Il en résulte que

$$E(X) = p \quad \text{et} \quad Var(X) = \sqrt{\frac{pq}{n}}.$$

Conséquences.

1. La réponse à la question que nous nous posions est oui : l'espérance de la fréquence d'échantillon est égale à la probabilité théorique d'apparition dans la population.
2. Lorsque la taille de l'échantillon augmente, la variance de F diminue, ce qui est logique : plus on a d'informations, plus il est probable que la proportion observée dans l'échantillon soit proche de la proportion de la population.

3.5.2 Distribution de la proportion d'échantillon dans le cas des grands échantillons

On sait que si $n \geq 30$, $np \geq 15$ et $nq \geq 15$, on peut approcher la loi binomiale par la loi normale de même espérance et de même écart-type. Donc F suit approximativement $\mathcal{N}(p, \sqrt{\frac{pq}{n}})$, et la variable $T = \dfrac{F - p}{\sqrt{\frac{pq}{n}}}$ suit alors approximativement la loi $\mathcal{N}(0, 1)$.

Exercice 16 *Selon une étude sur le comportement du consommateur, 25% d'entre eux sont influencés par la marque, lors de l'achat d'un bien. Si on interroge 100 consommateurs pris au hasard, quelle est la probabilité pour qu'au moins 35 d'entre eux se déclarent influencés par la marque ?*

Solution de l'exercice 16. *Influence de la marque.*

Appelons F la variable aléatoire : "proportion d'échantillon dans un échantillon de taille 100". Il s'agit ici de la proportion de consommateurs dans l'échantillon qui se déclarent influencés par la marque. On cherche à calculer $P(F > 0.35)$.

Il nous faut donc déterminer la loi de F. Or $np = 100 \times 0.25 = 25$ et $nq = 100 \times 0.75 = 75$. Ces deux quantités étant supérieures à 15, on peut considérer que F suit $\mathcal{N}(p, \sqrt{\frac{pq}{n}}) = \mathcal{N}(0.25, 0.0433)$.

On utilise la variable $T = \frac{F - 0.25}{0.0433}$ qui suit la loi $\mathcal{N}(0, 1)$. Il vient

$$P(F > 0.35) = P(T > 2.31) = 0.5 - P(0 < T < 2.31) = 0.5 - 0.4896 = 0.0104.$$

Conclusion. Il y a environ une chance sur 100 pour que plus de 35 consommateurs dans un 100 - échantillon se disent influencés par la marque lorsque l'ensemble de la population contient 25% de tels consommateurs.

En pratique, il est peu fréquent de connaître p : on doit plutôt l'estimer à partir d'un échantillon. Comment faire ? C'est ce que nous traiterons dans le prochain chapitre.