

Introduction à la statistique et à l'économétrie -  
IUP Finance

Chapitre 1: Échantillonnage et premiers pas en  
estimation

Christelle Dumas

## Contents

<b>1</b>	<b>Population et échantillon aléatoire</b>	<b>5</b>
1.1	De l'importance de l'échantillon et de son choix . . . . .	5
1.2	Population . . . . .	6
1.3	Retour sur des notions de probas . . . . .	7
1.3.1	Définitions - rappels . . . . .	7
1.3.2	Lois à deux dimensions . . . . .	9
1.3.3	Propriétés de l'espérance mathématique et de la variance dans le cas d'une combinaison de plusieurs variables aléatoires	11
1.4	Échantillon aléatoire . . . . .	12
1.5	Tirage avec ou sans remise? . . . . .	14
<b>2</b>	<b>Distribution d'échantillonnage d'une moyenne</b>	<b>14</b>
2.1	Définition et exemple . . . . .	15
2.2	Des statistiques échantillonnales comme variables aléatoires . . .	15
2.3	Moments d'une moyenne . . . . .	17
2.3.1	Population infinie . . . . .	17
2.3.2	Population de taille finie . . . . .	18
2.3.3	Application . . . . .	19
2.4	Loi d'une moyenne échantillonnale . . . . .	19
2.4.1	Pourquoi il est important de caractériser la loi . . . . .	19
2.4.2	Retour sur le théorème central limite . . . . .	20

2.4.3	Variance de la population connue ou inconnue? . . . . .	21
2.5	Calculs de probabilités sur des moyennes d'échantillon . . . . .	22
2.5.1	Événements associés à $\bar{X}$ . . . . .	22
2.5.2	Calcul du seuil de probabilité et application au test d'aptitudes. . . . .	22
<b>3</b>	<b>Distribution d'échantillonnage d'une variance</b>	<b>23</b>
3.1	Définitions et exemple . . . . .	23
3.2	Espérance d'une variance échantillonnale . . . . .	24
<b>4</b>	<b>Distribution d'échantillonnage d'une proportion</b>	<b>26</b>
4.1	Définition . . . . .	26
4.2	La proportion est une moyenne . . . . .	27
4.3	Espérance et variance d'une proportion empirique . . . . .	27
4.3.1	La fréquence est une binômiale . . . . .	27
4.3.2	La proportion est une moyenne . . . . .	28
4.4	Distribution de la proportion . . . . .	28
4.4.1	Approximation par une normale . . . . .	28
4.4.2	Approximation par une Poisson . . . . .	29

## Bibliographie éclairée

Voici quelques livres de stats, probas et économétrie qui pourront vous aider cette année:

- pour une présentation très intuitive mais un peu rapide: “Statistique”, Wonacott, T.H. & Wonacott, R.J., *Economica*.
- pour une présentation très simple et plus détaillée: “Probabilité statistique et techniques de régression”, Baillargeon, G. *Les Editions SMG* ou “Statistique appliquée à la psychologie”, Martin, L. et Baillargeon, G., *Les Editions SMG*.
- pour une présentation appliquée à la gestion et au business: “Statistics for business and economics”, Anderson, Sweeney et Williams, *Thomson*.
- pour la suite “Lire l'économétrie”, Behaghel, L., *Collection Repères*.

## Introduction générale

**Qu'est-ce que les statistiques?** La plupart des événements qui ont lieu peuvent être pensés comme le résultat de processus ou d'expériences qui ont une composante incertaine. Quand une pomme tombe d'un arbre, par exemple, il est difficile de savoir à l'avance où la pomme va exactement tomber. Dans certains cas, c'est parce que l'on ne dispose pas de l'information complète - i.e. si l'on connaissait le poids exact, la pression atmosphérique, le vent, les effets gravitationnels des autres objets environnants, etc. on pourrait calculer très précisément où la pomme tomberait. Bien entendu, même avec toutes ces informations il y a la possibilité d'une incertitude au niveau quantique... De toute façon, dans la vie réelle, on ne dispose que très rarement de l'information complète. Le temps qu'il fait aujourd'hui, le temps d'attente pour avoir le RER..., tous ces événements ont une composante déterministe (c'est-à-dire fixée, valide de façon permanente) et une composante incertaine, aléatoire. Ainsi, si vous répétiez votre départ de chez vous de ce matin, en partant exactement des mêmes conditions initiales, il est très probable que le RER ne soit pas arrivé exactement en même temps (pas à la même nanoseconde en tout cas), que la personne dans laquelle vous vous êtes cogné ne soit pas la même, etc. Les aspects similaires entre ces deux départs peuvent être considérés comme déterministes, tandis que les différences sont aléatoires.

L'objet des statistiques est de penser la plupart des événements en ces termes. Quand on observe la décision d'un individu de travailler par exemple, elle est gouvernée à la fois par des facteurs objectifs (nombre d'enfants, revenu du conjoint, qualifications, etc.) et d'autres facteurs que l'on ne saisit pas bien et qui peuvent être soit des facteurs que l'on n'observe pas (préférences) ou une vraie composante aléatoire (la personne s'est levée du pied gauche le matin). Sur la base d'un seul individu, il sera impossible de démêler la partie déterministe de la partie aléatoire; mais avec beaucoup d'individus, il sera possible de voir émerger certains schémas un peu systématiques.

Cependant, que les résultats de nombre de processus soient incertains ne signifie pas que l'on n'a rien à dire à propos de ces processus. Si l'on repense au pommier par jour de grand vent, on sera en mesure de dire que le plus probable est que la pomme tombe au pied de l'arbre, qu'il est possible qu'elle tombe à 5 m de l'arbre si elle se décroche au moment d'une rafale mais qu'il est très peu probable qu'elle tombe à 1km du pommier. C'est implicitement ce que vous avez étudié dans les cours précédents en listant les différentes lois. Bien que les résultats soient aléatoires, dans la plupart des cas, certains résultats sont plus probables que d'autres. Bien que cela semble évident, cela implique que l'on va être en mesure de faire de bons paris et de mauvais paris. Par exemple, je ne vais pas être en mesure de prédire parfaitement l'endroit où la pomme va tomber, mais je peux faire le pari qu'elle va tomber sous l'arbre. Ce n'est qu'un pari mais c'est le moins idiot parce que le plus probable, c'est-à-dire celui où j'ai le moins de chances de me tromper. La statistique est la science qui identifie la meilleure procédure pour prédire l'occurrence d'un événement et donc de

prendre les meilleures décisions dans un environnement incertain.

**Et l'estimation? et les tests?** Au sein de cette science, on peut distinguer différentes composantes: **l'estimation**, qui consiste à proposer une estimation, une prédiction et **les tests**. La théorie des tests permet d'utiliser un ensemble d'observations pour tester une théorie. Reprenons l'exemple du pommier: j'observe un pommier pendant plusieurs jours et note l'endroit où tombent les pommes. La plupart des pommes tombent sous le pommier mais quelques-unes tombent un peu plus loin. Avec ceci, je vais être en mesure de tester si, en moyenne, les pommes tombent en-dessous du pommier (ça a l'air stupide dit comme ça?). Plus précisément, je vais pouvoir dire quelle est ma marge d'erreur quand je dis que les pommes tombent en-dessous du pommier: 5%, 10%, 30%? Pourquoi est-ce important? et bien disons que c'est l'heure de la sieste et que j'aimerais bien me mettre sous le pommier parce qu'il y a de l'ombre, mais que j'ai peur de recevoir une pomme sur la tête pendant mon sommeil. Ceci va me permettre de savoir à quelle distance il faut que je me mette du pommier pour diminuer mes chances de recevoir une pomme. Peut-être à 2m pour ne risquer la chute qu'à 10% mais si ça fait trop, à quelle distance? 3m? 4m?

L'utilisation des statistiques dans la théorie des tests est la suivante: dans la mesure où on n'observe qu'une partie des pommes tombées et non pas toutes les chutes de pommes possible, l'information est partielle. Il faut donc prendre en compte l'aléatoire de l'estimation pour choisir où on va faire sa sieste. Sans entrer trop dans le détail sur les tests à ce stade, retenons que nous avons une hypothèse sur la façon dont le monde fonctionne (les pommes tombent sous les pommiers) et nous voudrions savoir si notre hypothèse est correcte ou non. Les applications de ceci sont multiples: vous faites un sondage pour déterminer le vainqueur de l'élection présidentielle et vous trouvez que Mme R. gagne contre M. S. à 75%. Que croyez-vous pour la vraie élection (ie sondage avec la population totale)? et si c'était 50.5%? est-ce que ça dépend du nombre de personnes sondées? Vous voulez savoir si les burgers font grossir et vous pesez les mangeurs de burgers et les autres. Et vous trouvez une différence de 1kg. Qu'est-ce que ça veut dire?

**L'objet de l'économétrie** Décrire un échantillon et donc une population à l'aide de moyennes, de fréquences ou de variance est un objet en soi mais reste très basique. Quid si l'on s'intéresse à l'effet d'une augmentation du prix des cigarettes sur la consommation des individus, par exemple? Or c'est typiquement ce que le gouvernement aimerait savoir avant de mettre en place une nouvelle taxe. Pour cela, on regarde les fluctuations des deux variables clefs: prix de la cigarette et consommation et on en dérive une estimation de l'effet de l'un sur l'autre. Cependant, un nombre important de paramètres peuvent varier en même temps (par exemple, l'existence de campagnes anti-tabac) et on ne voudrait pas imputer l'effet des campagnes anti-tabac à l'augmentation du prix de la cigarette. La force de l'économétrie est d'identifier un effet très bien

défini en prenant en compte les autres changements dans l'environnement. Ce sera l'objet de la fin de ce cours. Les champs décrits ne sont pas disjoints les uns des autres mais s'emboîtent donc, par exemple, on sera en mesure d'appliquer la théorie des tests à l'économétrie et à dire si l'impact d'une hausse des prix des cigarettes a un effet significativement différent de zéro sur la consommation ou non.

## 1 Population et échantillon aléatoire

### 1.1 De l'importance de l'échantillon et de son choix

"You don't have to eat the whole ox to know that it is tough."

Samuel Johnson

Les statistiques sont utiles pour se contenter d'une portion de l'information, ie: ne pas interroger toutes les personnes (recensement) mais seulement un sous-ensemble.

- collecter des données exhaustives est très coûteux (même le recensement n'est plus exhaustif); arbitrage entre le nombre d'individus enquêtés et la teneur de l'enquête.
- inconvénient: risque que les personnes enquêtées ne soient pas représentatives de la population totale.
- prise en compte du fait qu'il ne s'agit que d'un échantillon et pas de la population totale.

Échantillonner plutôt que recenser pose une double question:

- comment choisir son échantillon? → théorie des sondages.
- que peut-on dire sur la totalité de la population lorsque l'on utilise seulement une partie de cette population? → inférence statistique: méthode visant à retrouver les caractéristiques d'une population à partir des renseignements contenus dans un échantillon.

#### **La représentativité et le sondage aléatoire: des exemples d'échantillon problématiques**

- L'énigme de Flipper le dauphin ou "mais pourquoi sont-ils si gentils?"
- Sondage par téléphone: la consommation de burgers chez les ménagères de moins de 50 ans ou l'importance de définir la population cible.

## 1.2 Population

Soit un ensemble comprenant un certain nombre d'objets à étudier. Cet ensemble est appelé **la population** ( $\mathcal{P}$ ). Les éléments de cet ensemble sont appelés les **individus**. Le nombre d'individus ( $N$ ) représente **la taille ou la dimension** de la population. La taille de la population peut-être variable et nous supposons, au moins dans un premier temps, que la taille de la population est importante ( $N > 30$ ) et finie.

Le point sur lequel les individus diffèrent est appelé le **caractère**<sup>1</sup> ( $X$ ). Ce caractère peut-être quantitatif (*mesurable*) ou qualitatif (*non mesurable*).

**Définition 1** *La population est l'ensemble de toutes les valeurs observables  $x_j \forall j = 1 \dots N$  d'une variable  $X$  décrivant un caractère.*

Plus concrètement (et pour les propos qui nous concernent), **la population est l'ensemble des individus à étudier à partir duquel est extrait l'échantillon.**

Exemple: l'ensemble des élèves de sexe féminin d'une classe de manequinat constitue une population (d'individus). L'ensemble des tailles de ces femmes est aussi appelé population. Le tableau 1 indique la distribution de cette population.

Table 1. Distribution de la population

Taille	Fréquence	Fréquence relative $p(x)$
1m66	2	0.02
1m69	7	0.07
1m72	24	0.24
1m75	37	0.37
1m78	23	0.23
1m81	6	0.06
1m84	1	0.01
Total	100	1.00

La *taille* de cette population est  $N = 100$ . Le *caractère* auquel nous nous intéressons est un caractère quantitatif (donc mesurable), *i.e.* la taille des jeunes filles mannequins. Ici c'est un caractère discret (le caractère taille prend des modalités discrètes, en nombre fini). La moyenne et la variance (écart-type) de la population sont respectivement données par  $\mu = 174.8$  cm et  $\sigma^2 = 11.30$  ( $\sigma = 3.36$  cm). Il s'agit de la moyenne et de la variance observées du caractère "taille" observé sur la population.

<sup>1</sup>Nous nous limiterons au caractère statistique unidimensionnel.

On pourrait donner un exemple similaire pour illustrer un caractère qualitatif: la couleur des yeux des jeunes femmes, par exemple.

**Question qu'on se pose** : “quelle est la probabilité qu'un individu tiré au hasard dans la population soit de 175 cm?” : on est maintenant du côté probabilité : à cette expérience, est associée une variable aléatoire  $X$  (taille de l'individu tiré au hasard), déterminée par les modalités qu'elle peut prendre et la probabilité d'apparition de ces valeurs. Il s'agit de déterminer la loi de probabilité de cette variable aléatoire  $X$ . Comme tous les individus  $i = 1$  à 100 ont la même probabilité d'être choisis, la probabilité pour un individu quelconque est  $1/100$ . Donc on recherche dans la population, pour calculer  $\Pr(X_i = 175)$ , le nombre de cas favorables (37 individus font 1.75m) rapporté au nombre de cas possibles (100). La probabilité cherchée est donc  $\Pr(X = 175) = 37/100 = 0.37$ , ie la fréquence relative.

**La colonne  $p(x)$  peut donc être vue comme la distribution empirique du caractère "taille" dans la population, et comme la distribution de probabilité de la variable aléatoire  $X$  "taille". On a là une idée importante de l'échantillonnage : chaque individu dans la population a la même distribution de probabilité que la population, qui correspond à la distribution empirique du caractère.**

On peut donc mettre en regard les concepts probabilistes et les concepts statistiques de la façon indiquée dans le tableau 2.

Table 2. Concepts probabilistes et statistiques

Notions probabilistes	Notions statistiques
Probabilité d'un événement	Fréquence relative
Variable aléatoire	Variable statistique
Loi de probabilité	Distribution statistique (empirique)
Espérance mathématique d'une v.a.	Moyenne arithmétique d'une variable statistique
Variance d'une v.a.	Variance d'une variable statistique

## 1.3 Retour sur des notions de probas

### 1.3.1 Définitions - rappels

**Définition 2** *Variable aléatoire*: Si à chaque résultat d'une expérience aléatoire (comme un tirage), on fait correspondre une valeur numérique, nous disons alors que l'on a une variable aléatoire.

**Définition 3** *Loi de probabilité:* Associer à chacune des valeurs possibles de la variable aléatoire la probabilité qui lui correspond c'est définir la loi (ou distribution) de probabilité de la v.a.

On représente différemment cette loi de probabilité selon que la variable aléatoire est discrète (ex: tirage d'un dé à 6 faces, nombre limité de valeurs possibles) ou continue (ex: temps d'attente du RER, nombre illimité de valeurs possibles). Si la variable est discrète, on utilise la fréquence relative, si elle est continue, on utilise la densité.

**Définition 4** *Densité d'une v.a.:* la densité d'une v.a.  $X$  en  $x$  est:

$$f(x) = \lim_{dx \rightarrow 0} P(x \leq X \leq x + dx).$$

**Définition 5** *Fonction de répartition ou cumulative:* la fonction de répartition est la probabilité cumulée des valeurs de  $X$  jusqu'à  $x_i$ :

$$F(x_i) = P(X \leq x_i).$$

**Définition 6** *Espérance mathématique:* Soit  $X$  une variable aléatoire qui prend des valeurs sur un support non discret  $[a; b]$  et dont la loi de probabilité est représentée par la densité  $f$ . L'espérance mathématique de  $X$ , notée  $E(X)$  s'écrit:

$$E(X) = \int_a^b x \cdot f(x) \cdot dx.$$

Pour une variable aléatoire discrète qui prendrait les valeurs  $x_1, \dots, x_n$ , ceci se réécrit:

$$E(X) = \sum_{i=1}^n x_i \cdot P(X = x_i)$$

**Définition 7** *Variance et écart-type:* La dispersion des valeurs de la variable aléatoire est obtenue en calculant l'espérance des carrés des écarts de ces valeurs à l'espérance mathématique, c'est-à-dire la valeur moyenne des carrés  $(x_i - E(X))^2$ :

Pour une variable aléatoire continue, ceci s'écrit:

$$V(X) = \int_a^b (x - E(X))^2 f(x) dx$$

et pour une variable discrète:

$$V(X) = \sum_{i=1}^n (x_i - E(X))^2 P(X = x_i).$$

La racine carrée de la variance se nomme l'écart-type et a la même unité que celle de la variable aléatoire.

**Propriété 8** de l'espérance et de la variance

$$\begin{aligned}E(aX + c) &= aE(X) + c \\E(X + Y) &= E(X) + E(Y) \\V(aX + c) &= a^2V(X)\end{aligned}$$

où  $a$  et  $c$  sont des constantes et  $X$  et  $Y$  des variables aléatoires.

### 1.3.2 Lois à deux dimensions

**Définition 9** Loi conjointe de deux variables aléatoires discrètes: Soient  $X$  et  $Y$  deux variables aléatoires discrètes dont l'ensemble des valeurs possibles sont respectivement  $X = x_1, x_2, \dots, x_m$  et  $Y = y_1, y_2, \dots, y_n$ . Associer à chacune des valeurs possibles du couple  $(X, Y)$  la probabilité  $f(x_i, y_j)$  que  $X$  prenne la valeur  $x_i$  et  $Y$  la valeur  $y_j$ , c'est définir la loi conjointe des variables aléatoires  $X$  et  $Y$ :

$$f(x_i, y_j) = P(X = x_i, Y = y_j).$$

Le couple  $(X, Y)$  s'appelle également variable aléatoire à 2 dimensions et peut prendre  $m \cdot n$  valeurs.

**Définition 10** Lois marginales: Soit la variables aléatoire  $(X, Y)$  à deux dimensions admettant comme loi conjointe  $f(x_i, y_j)$ . Alors, les lois marginales de  $X$  et de  $Y$  sont définies respectivement par:

$$\begin{aligned}f(x_i) &= P(X = x_i) = \sum_{j=1}^n f(x_i, y_j), \quad \forall i = 1, \dots, m \\f(y_j) &= P(Y = y_j) = \sum_{i=1}^m f(x_i, y_j), \quad \forall j = 1, \dots, n\end{aligned}$$

**Définition 11** Lois conditionnelles: Soit la variable aléatoire  $(X, Y)$  à deux dimensions admettant comme loi conjointe  $f(x_i, y_j)$  et comme lois marginales  $f(x_i)$  et  $f(y_j)$ . Supposons que la probabilité que  $X$  prenne la valeur  $x_i$  ne soit pas nulle, alors la probabilité conditionnelle de  $Y = y_j$  sachant que  $X = x_i$  s'est réalisé est définie par:

$$f(y_j|x_i) = \frac{f(x_i, y_j)}{f(x_i)}$$

Les probabilités  $f(y_j|x_i)$  associées aux différentes valeurs possibles  $y_j$  de  $Y$  constituent la loi conditionnelle de  $Y$ .

**Définition 12** Indépendance: Soit la variable aléatoire  $(X, Y)$  à deux dimensions admettant comme loi conjointe  $f(x_i, y_j)$  et comme lois marginales  $f(x_i)$

et  $f(y_j)$ . On dit que les variables aléatoires sont indépendantes si et seulement si les probabilités conjointes sont égales au produit des probabilités marginales:

$$f(x_i, y_j) = f(x_i) \cdot f(y_j)$$

pour toutes les valeurs  $(x_i, y_j)$ .

Remarque: on peut définir de façon équivalente l'indépendance de deux variables aléatoires. Les variables aléatoires  $X$  et  $Y$  sont indépendantes si la loi conditionnelle de  $X$  pour toute valeur de  $Y$  est identique à la loi marginale de  $X$  et si la loi conditionnelle de  $Y$ , pour toute valeur de  $X$ , est identique à la loi marginale de  $Y$ :

$$f(x_i|y_j) = f(x_i)$$

$$f(y_j|x_i) = f(y_j)$$

**Application** Catégorie salariale et ancienneté. Une entreprise a classé ses 300 employés selon les catégories salariales suivantes indiquées dans la table 3; on définit  $X$  comme étant la catégorie salariale dans laquelle peut se situer un employé pris au hasard.

Table 3. Distribution de la catégorie salariale

Valeur de $X$	Catégorie salariale	Nombre d'employés	$P(X = x_i)$
1	25000 euros et moins	75	0.25
2	Entre 25000 et 35000 euros	120	0.40
3	35000 euros et plus	105	0.35

Par ailleurs, on a également noté un autre caractère, l'ancienneté de chaque employé selon les catégories salariales (table 4). On notera  $Y$  la classe d'ancienneté

Table 4. Répartition de l'ancienneté selon les catégories salariales

Catégorie salariale	Ancienneté		
	4 ans et moins	Entre 4 et 10 ans	10 ans et plus
25000 euros et moins	57	15	3
Entre 25000 et 35000 euros	12	81	27
35000 euros et plus	0	45	60

à laquelle appartient l'employé. La répartition de cette variable aléatoire est consignée dans le tableau 5.

Table 5. Distribution de l'ancienneté

$y_j$	Ancienneté	Nombre d'employés	$P(Y = y_j)$
0	4 ans et moins	69	0.23
1	Entre 4 et 10 ans	141	0.47
2	10 ans ou plus	90	0.30

- Distribution conjointe de  $X$  et de  $Y$ ?
- Représentation graphique?
- Loi marginale de  $X$ ? de  $Y$ ?
- Lois conditionnelles?
- Indépendance?

### 1.3.3 Propriétés de l'espérance mathématique et de la variance dans le cas d'une combinaison de plusieurs variables aléatoires

**Propriété 13** Si  $X$  et  $Y$  sont deux v.a. indépendantes, alors:

$$\begin{aligned} V(X + Y) &= V(X) + V(Y) \\ E(X \cdot Y) &= E(X) \cdot E(Y) \end{aligned}$$

**Définition 14** Covariance de deux v.a. Soient  $X$  et  $Y$  deux variables aléatoires. La covariance de  $X$  et de  $Y$  est l'espérance mathématique du produit des écarts de  $X$  et de  $Y$  à leurs espérances mathématiques respectives:

$$\begin{aligned} Cov(X, Y) &= E[(X - E(X)) \cdot (Y - E(Y))] \\ &= E(X \cdot Y) - E(X) \cdot E(Y) \end{aligned}$$

Si les deux variables aléatoires sont indépendantes, alors

$$Cov(X, Y) = E(X) \cdot E(Y) - E(X) \cdot E(Y) = 0.$$

**Définition 15** Corrélation de deux v.a. On définit le coefficient de corrélation  $\rho$  entre  $X$  et  $Y$  par le rapport:

$$\rho = \frac{Cov(X, Y)}{\sqrt{V(X) \cdot V(Y)}}$$

Le coefficient de corrélation est une mesure de l'intensité de la liaison linéaire entre 2 variables aléatoires. On peut montrer que  $\rho$  peut varier entre -1 (corrélacion parfaite négative) et +1 (corrélacion parfaite positive). Si  $\rho = 0$ , on dit que les variables aléatoires sont non-corrélées (absence de liaison linéaire).

Remarques importantes:

- Si les variables aléatoires  $X$  et  $Y$  sont indépendantes, alors  $Cov(X, Y) = 0$  et  $\rho = 0$ . À l'inverse, si deux variables aléatoires ne sont pas corrélées, alors elles ne sont pas nécessairement indépendantes (possibilité d'une liaison autre que linéaire).
- Pour deux variables aléatoires  $X$  et  $Y$  non nécessairement indépendantes,

$$V(X + Y) = V(X) + 2Cov(X, Y) + V(Y).$$

## 1.4 Échantillon aléatoire

Comme on l'a vu, un sondage doit reposer sur un choix adéquat d'échantillon. Il s'agit donc de savoir comment constituer l'échantillon pour qu'il soit représentatif de la population. N'oubliez pas que, *in fine*, l'objectif est de faire de l'inférence statistique, c'est-à-dire de dire quelque chose sur la population totale à partir de l'échantillon.

**Définition 16** *Un échantillon de taille  $n$   $(X_1, \dots, X_n)_{1 \leq i \leq n}$  est l'ensemble de  $n$  v.a.r. correspondant à  $n$  tirages effectués dans une population de référence  $\mathcal{P}$  (de taille  $N$  ou infinie). On note  $x$  la réalisation de la v.a.r  $X$ .*

Ou, plus simplement, un échantillon tiré de la population est un sous-ensemble de cette population. On appelle  $n$  la dimension ou la taille de l'échantillon. On dit que l'échantillon est **représentatif** si les résultats de l'analyse conduite sur cet échantillon peuvent être étendus à l'ensemble de la population.

Un échantillon **aléatoire** est un échantillon pour lequel on a choisi au hasard le sous-ensemble d'individus (on verra diverses techniques de tirage aléatoire plus loin). L'échantillonnage aléatoire est une technique qui permet d'obtenir un échantillon représentatif de la population.

**Comment tirer aléatoirement dans une population?** Diverses options:

- **Le sondage aléatoire simple ou élémentaire** cette méthode consiste à tirer  $n$  individus de la population en donnant à chaque individu de la même probabilité d'être désigné;
- **Le sondage systématique** cette méthode consiste à choisir les individus à des intervalles fixes (temps, espace,...) selon un pas déterminé à l'avance; (exemple : on interroge tous les mois les même groupe d'individus, on interroge tous les habitants d'un quartier dans l'adresse est au n° 5,...

- **Le sondage stratifié** cette méthode consiste à subdiviser la population en groupes relativement homogènes, *i.e.* des strates; (on interroge par CSP, etc...)
- **Le sondage par grappes** les méthodes précédentes utilisent tous les individus composant la population de référence mais il peut être difficile d'établir la liste complète de tous les individus de la population. Dans ce cas la population est partagée en groupes (*grappes*). Par exemple, un ménage, un ensemble de personnes habitant un même logement etc...

Dans la suite de ce cours, nous nous limiterons aux échantillons aléatoires simples <sup>2</sup>.

Comment fait-on un tirage aléatoire simple? Première option: inscrire toutes les valeurs de la population sur un petit papier, mettre tous les papiers dans un chapeau puis tirer des petits papiers. Seconde option: numéroter tous les individus puis choisir aléatoirement des numéros (en suivant une table par exemple) et inscrire la valeur pour ces individus. Les deux procédés sont équivalents.

Exemple de tirage aléatoire de 10 individus dans les 100 élèves mannequins:

$$x_1 = 175, \quad x_2 = 172, \quad x_3 = 169, \quad x_4 = 181, \quad x_5 = 172, \\ x_6 = 172, \quad x_7 = 178, \quad x_8 = 175, \quad x_9 = 175, \quad x_{10} = 175.$$

La distribution de l'échantillon est indiquée dans le tableau 6.

Table 6. Distribution de l'échantillon

Taille	Fréquence	Fréquence relative
1m66	0	0.00
1m69	1	0.10
1m72	3	0.30
1m75	3	0.30
1m78	2	0.20
1m81	1	0.10
1m84	0	0.00
Total	10	1.00

**À retenir** Chaque  $X_i$  peut être considéré comme une v.a.r. dont la loi est donnée par la répartition des valeurs de  $X$  dans la population.

<sup>2</sup>Pour la suite du cours, nous désignerons les EAS simplement par échantillons aléatoires.

Intuition: avant le tirage, la valeur de la variable est inconnue, d'où l'aléatoire.  $X_1$ , la première valeur tirée est une variable aléatoire. Une fois le tirage effectué, on connaît sa valeur, notée  $x_1$ , égale par exemple à 1m75.

## 1.5 Tirage avec ou sans remise?

Faut-il remettre dans le chapeau les observations qui ont déjà été tirées?

- Contre: on préfère généralement ne pas utiliser 2 fois la même observation (l'information additionnelle lors de la deuxième utilisation est nulle);
- Pour: si l'on retire certaines observations, la distribution dans laquelle on tire les observations suivantes a changé et il faut le prendre en compte. Notamment, la loi de  $X_2$  dépend de la réalisation  $x_1$  et du coup les deux tirages ne sont pas indépendants.

Heureusement, si le nombre d'observations est grand, 1) il est très peu probable de tirer deux fois la même observation 2) la distribution change très peu si on enlève seulement une observation. Donc: quand le nombre d'observations est grand, on considère que la loi ne change pas et que les observations sont indépendantes. Quand le nombre d'observations est petit, on effectue un tirage sans remise et on doit prendre en compte le fait que la loi change.

**Proposition 17** *Un échantillon aléatoire très simple (EATS) est un échantillon dont les  $n$  observations  $X_1, X_2, \dots, X_n$  sont indépendantes. À chaque tirage est associé une variable aléatoire et la loi de chaque v.a.  $X_i$  est la même que celle de la population. Chaque individu a alors la même espérance  $m$  et le même écart-type  $\sigma$  que la population.*

## 2 Distribution d'échantillonnage d'une moyenne

Dans l'exemple des femmes mannequins, nous connaissons la distribution exacte du caractère taille, et donc notamment ses moments (moyenne, variance, fréquence d'une modalité...). Mais si l'on ne fait pas un recensement, on ne connaît pas la distribution exacte. **Le but de la théorie de l'échantillonnage est de déterminer, à partir de l'observation sur échantillon, des procédures d'induction qui permettent d'aller de l'échantillon vers la population.** Imaginons que l'on cherche à connaître la moyenne et la variance des tailles d'élèves mannequins en utilisant uniquement un échantillon aléatoire (représentatif). On cherche à calculer un paramètre qu'on appellera  $\theta$  (ici moyenne, fréquence, variance) inconnu. On appelle ça "estimer"  $\theta$ . Il faudra aussi se demander quelle confiance on peut avoir en des résultats obtenus seulement sur un échantillon.

## 2.1 Définition et exemple

**Définition 18** *La moyenne d'échantillon est définie comme suit:*

$$\overline{X}_n \equiv \frac{\sum_{i=1}^n X_i}{n}$$

On note  $\overline{x}_n \equiv \frac{\sum_{i=1}^n x_i}{n}$  la réalisation de la v.a.r.  $\overline{X}_n$ . Il arrive que l'on note  $\overline{X}$  la moyenne échantillonnale en omettant l'indice. .

Application au tirage des 10 élèves mannequins: la moyenne de l'échantillon des 10 mannequins est  $\overline{x}_{10} = 174.7$

**Une question que l'on se pose immédiatement mais à laquelle on répondra plus tard:** est-ce que l'on considère que la moyenne de l'échantillon est un bon estimateur de la moyenne de la population? Dans ce cas, c'est très proche, mais si par manque de chance nous avons tiré uniquement les élèves les plus petites, on arriverait à 168.7 cm. Le problème, c'est que l'on veut être en mesure de déterminer si une procédure d'estimation est bonne ou non indépendamment du tirage que l'on effectue, puisque l'on ne connaît pas la vraie valeur. Il ne s'agit pas de comparer 174.7 ou 168.7 à 174.8 mais de savoir si la procédure qui consiste à prendre la moyenne d'échantillon comme estimateur de la moyenne de la distribution est acceptable. Il faudra donc définir des critères.

Avant cela, puisque la moyenne d'échantillonnage est une variable aléatoire, on peut se demander quelle est sa distribution, ce qui sera un premier pas vers la question de l'estimation. Mais encore en amont, je vais tâcher de vous convaincre que la moyenne d'un échantillon est une variable aléatoire.

**Définition 19** *Une statistique est une variable aléatoire dont les valeurs ont été obtenus à partir d'échantillons de même taille extraits de la même population. Ces valeurs sont, par exemple, la moyenne d'échantillon, sa variance ou la proportion, etc. Cette variable aléatoire est caractérisée par une loi de probabilité qu'on appelle **loi (ou distribution) d'échantillonnage**. Ainsi, une distribution d'échantillonnage est la loi de probabilité d'une statistique.*

*Une statistique  $\hat{\theta}_n = f(X_1, \dots, X_n)$  est une v.a.r. calculée à partir des v.a.r. de l'échantillon, suivant une règle définie à l'avance.*

*La distribution d'échantillonnage de la statistique  $\hat{\theta}_n$  est la loi (discrète ou continue) de la v.a.r.  $\hat{\theta}_n$ .*

## 2.2 Des statistiques échantillonnales comme variables aléatoires

Prenons une population de 5 individus à qui on a alloué un certain montant pour un voyage d'études: Luc reçoit 600 kopecks, André 150, Viviane 300, Pierre 600

et Robert 150. La moyenne de la population est  $\mu = 360$ , la variance est:

$$\sigma^2 = V(X) = \sum x^2 f(x) - \mu^2 = 171000 - (360)^2 = 41400.$$

La proportion des montants supérieurs à 200 kopecks est  $p = 0.60$ .

Tirons, sans remise, les 10 échantillons possibles de taille 3. Et pour chacun calculons la moyenne échantillonnale  $\bar{x}$ ; les résultats sont consignés dans le tableau 7.

Table 7. Moyenne pour l'ensemble des échantillons

Echantillon n°	$(i, j, k)$	Résultat de l'échantillonnage	$\bar{x}$
1	(1,2,3)	(600,150,300)	350
2	(1,2,4)	(600,150,600)	450
3	(1,2,5)	(600,150,150)	300
4	(1,3,4)	(600,300,600)	500
5	(1,3,5)	(600,300,150)	350
6	(1,4,5)	(600,600,150)	450
7	(2,3,4)	(150,300,600)	350
8	(2,3,5)	(150,300,150)	200
9	(2,4,5)	(150,600,150)	300
10	(3,4,5)	(300,600,150)	350

Il apparaît donc que la statistique  $\bar{X}$  est une variable aléatoire dont la réalisation dépend de l'échantillon tiré. Cette statistique a donc une distribution, qui correspond aux différentes valeurs que la statistique peut prendre lorsque l'on décrit l'ensemble des échantillons.

La statistique  $\bar{X}$  associe à chaque échantillon aléatoire (ici, de taille 3) sa moyenne  $\bar{x}$ . Le tableau 8 décrit sa distribution:

Table 8. Distribution de la statistique  $\bar{X}$

$\bar{x}$	$Pr(\bar{X} = \bar{x})$
200	1/10
300	2/10
350	4/10
450	2/10
500	1/10

À partir de cette distribution, on peut calculer l'espérance et la variance de  $\bar{X}$ . Dans ce cas précis, on obtient:  $E(\bar{X}) = 360$  et  $V(\bar{X}) = 6900$ .

Normalement, à ce stade, vous devez être convaincus que les statistiques telles que la moyenne échantillonnale sont des variables aléatoires (sinon, relisez la précédente section!!). Dans les sections suivantes, nous indiquerons à quelle loi de probabilité obéit une statistique et comment son espérance et sa variance sont reliées à celles de la population.

## 2.3 Moments d'une moyenne

Comme on vient de le voir, la moyenne d'un échantillon n'est pas nécessairement égale à la moyenne sur l'ensemble de la population<sup>3</sup>. Or il paraît assez intuitif d'utiliser la moyenne sur un échantillon pour estimer la moyenne sur la population totale. Mais pour cela, il faut prendre en compte les "erreurs d'échantillonnage", qui ne sont pas de vraies erreurs, mais simplement le fait que l'on ne peut pas espérer avoir une mesure exacte en se contentant d'une partie de la population.

On peut quand même noter que, sur l'exemple précédent,

$$E(X) = 360 = E(\bar{X})$$

ce qui est l'indication que, en général, si on veut "deviner" une moyenne, on peut utiliser la moyenne d'un échantillon. Par ailleurs, on remarquera que la dispersion de la moyenne est bien plus faible que celle des observations: sur l'exemple précédent, la variance de la population vaut  $\sigma^2 = 41400$  tandis que la variance de la moyenne vaut  $V(\bar{X}) = 6900$ . C'est intuitif puisque quand on moyenne, en général, une valeur élevée vient contre-balancer une valeur faible. L'objet de cette section est de formaliser l'ensemble de ces résultats et de mettre en évidence des propriétés de la moyenne.

### 2.3.1 Population infinie

**Proposition 20** *Si on prélève un échantillon aléatoire de taille  $n$  dans une population infinie dont le caractère mesurable  $X$  est régi par une loi de probabilité d'espérance  $E(X) = \mu$  et de variance  $V(X) = \sigma^2$ , alors:*

$$\text{l'espérance de la moyenne de } X \text{ est: } E(\bar{X}) = E(X) = \mu$$

$$\text{la variance de la moyenne de } X \text{ est: } V(\bar{X}) = \frac{\sigma^2}{n}$$

Démonstration pour l'espérance:

$$E\bar{X} = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n EX_i = \frac{1}{n}\sum_{i=1}^n \mu = \frac{1}{n}n\mu = \mu$$

---

<sup>3</sup>Sauf si l'échantillon est la population totale, bien sûr!

Démonstration pour la variance:

$$V(\bar{X}) = V\left(\frac{1}{n} \sum_i X_i\right) = \frac{1}{n^2} V\left(\sum_i X_i\right) = \frac{1}{n^2} \sum_i V(X_i) = \frac{\sigma^2}{n}$$

**Attention** : ne pas confondre  $\sigma^2$  et  $V(\bar{X}_n)$  qui portent tous deux le nom de variance.  $\sigma^2$  est la variance du caractère dans la population alors que  $V(\bar{X}_n)$  est la variance de la moyenne de l'échantillon. Ce résultat signifie que la dispersion de la statistique  $\bar{X}_n$ , ou la dispersion d'une infinité de réalisations  $\bar{x}_n$ ,  $V(\bar{X}_n)$ , autour de son espérance mathématique  $\mu$  (la vraie moyenne dans la population), est

- d'autant plus forte que la dispersion du caractère  $X$  dans la population est forte ( $\sigma^2$  au numérateur);
- et d'autant plus faible que la taille de l'échantillon est élevée ( $n$  au dénominateur);

Cette caractéristique de l'écart entre  $\bar{X}$  et  $\mu$  (la cible, *i.e.* la vraie valeur du paramètre) représente l'erreur d'estimation ou Standard-Error (SE) ou encore Ecart type d'échantillon. Nous avons donc :

$$SE = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (1)$$

### 2.3.2 Population de taille finie

Si la population est de taille finie  $N$ , deux cas se présentent à nous:

1. si l'échantillonnage se fait **avec remise**, alors la proposition reste valide.
2. si l'échantillonnage se fait **sans remise** dans une population de taille  $N$ , alors, pour la formule de la variance, on doit apporter un terme de correction pour prendre en compte la réduction de la taille de la population au fur et à mesure des tirages. La formule de la variance devient alors:

$$V(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \simeq \frac{\sigma^2}{n} \cdot \left(1 - \frac{n}{N}\right).$$

En pratique, on considère la propriété valide tant que la taille de l'échantillon ( $n$ ) reste faible par rapport à la taille de la population ( $N$ ), c'est-à-dire taux de sondage inférieur à 5% ou 10%. Lorsque le taux de sondage est supérieur à 10%, il faut appliquer la formule de la variance corrigée indiquée plus haut.

### 2.3.3 Application

Revenons à notre premier exemple, la population des jeunes étudiantes en manequinat est telle que :

$$E(X) = \mu = 174.8 \text{ et } V(X) = \sigma^2 = 11.30 \text{ (} \sigma = 3.36 \text{)}$$

#### Question

- (a) Si l'on tire plusieurs échantillons de taille  $n = 4$  et si l'on calcule à chaque fois la moyenne de l'échantillon  $\bar{X}$ , comment fluctuent ces moyennes de l'échantillon ?
- (b) Dans l'hypothèse où l'on quadruple la taille de l'échantillon soit  $n = 16$ , comment fluctuent les moyennes de l'échantillon ?

Réponse (a) La taille de l'échantillon est  $n = 4$ . La moyenne des échantillons (la moyenne de la distribution d'échantillonnage) est simplement égale à :

$$E(\bar{X}) = E(X) = \mu = 174.8$$

et la variance des moyennes (ou écart d'échantillon) vérifie :

$$V(\bar{X}) = \frac{\sigma^2}{n} = \frac{11.30}{4} = 2.825 \text{ soit } SE = 1.68$$

Ainsi les nombreuses valeur de la moyenne de l'échantillon  $\bar{X}$  varient autour de l'objectif de 174.8 cm avec une erreur d'estimation de plus ou moins 1.68 cm.

Réponse (b) La taille de l'échantillon est  $n = 16$ . La moyenne des échantillons est inchangée mais avec un écart-type différent. Il vient :

$$V(\bar{X}) = \frac{\sigma^2}{n} = \frac{11.30}{16} = 0.706 \text{ soit } SE = 0.84$$

Ainsi, **un quadruplement de la taille de l'échantillon entraîne un doublement de la précision.**

## 2.4 Loi d'une moyenne échantillonnale

### 2.4.1 Pourquoi il est important de caractériser la loi

La moyenne et la variance ne caractérisent pas entièrement la loi d'une variable aléatoire. Pour cela, il faut la distribution (ou loi) de la variable aléatoire. Notamment, on ne peut pas faire de tests sans connaître les lois des statistiques.

Inclure graphes

### 2.4.2 Retour sur le théorème central limite

Ainsi, nous verrons que si nous voulons faire des tests sur des moyennes, nous avons besoin de caractériser complètement les fluctuations d'échantillonnage de  $\bar{X}$ , c'est-à-dire décrire sa distribution.

Pour connaître exactement la loi de  $\bar{X}$ , nous avons besoin de la distribution de la population (comme dans l'exemple), or c'est rarement possible. Toutefois, un théorème essentiel en statistique va nous permettre de contourner cette difficulté.

#### **Théorème 21** *Théorème central limite*

*Si des échantillons aléatoires de taille  $n$  sont prélevés d'une population infinie dont les éléments possèdent un caractère mesurable  $X$  (peu importe la distribution de la variable aléatoire  $X$ ), de moyenne  $E(X) = \mu$  et de variance  $V(X) = \sigma^2$ , alors la distribution d'échantillonnage de la variable aléatoire  $\bar{X}$  tend à se rapprocher d'une loi normale de moyenne  $E(\bar{X}) = \mu$  et de variance  $V(\bar{X}) = \frac{\sigma^2}{n}$  et ce, d'autant plus que la taille de l'échantillon est grande.*

Enoncé formalisé:

Soit  $(X_i)_{i \in \mathbb{N}^*}$  suite de v.a.r. i.i.d. /  $\mathbb{E}X_i = \mu_X < +\infty$  et  $\mathbb{V}X_i = \sigma_X^2 < +\infty$ ;

$$\bar{X}_n \equiv \frac{1}{n} \sum_{i=1}^n X_i \quad \forall n \in \mathbb{N}^*.$$

alors  $\frac{\sqrt{n} \cdot (\bar{X}_n - \mu_X)}{\sigma_X} \xrightarrow{\mathcal{L}} \mathcal{N}(0; 1)$ .

Par abus de langage, on dit que " $\bar{X}_n$  suit approximativement une  $\mathcal{N}\left(\mu_X; \frac{\sigma_X^2}{n}\right)$  pour  $n$  grand":  $\bar{X}_n \underset{n \text{ grand}}{\rightsquigarrow} \mathcal{N}\left(\mu_X; \frac{\sigma_X^2}{n}\right)$ :

**$n$  grand?** À partir de quand considère-t-on que  $n$  est grand? en général, lorsque  $n \geq 30$ , on considère que l'on peut appliquer le théorème. Ceci dit, on a la même propriété lorsque  $n$  est petit mais que les observations  $X$  sont distribuées normalement.

**Conditions d'applications du théorème** Ce théorème est excessivement utile dans la pratique puisqu'il n'impose aucune restriction sur la distribution des observations de la population. Tant que la moyenne et la variance de la population existent, la distribution de la moyenne  $\bar{X}$  approche celle d'une normale à mesure que la taille d'échantillon augmente.

Si la forme de la distribution de la population est pratiquement symétrique, on peut se contenter de 15 observations pour considérer que l'on est proche de la loi normale.

Si la distribution de la population est normale, alors la distribution de la moyenne échantillonnale est normale quelque soit la taille de l'échantillon.

### 2.4.3 Variance de la population connue ou inconnue?

Jusqu'ici, on a fait comme si la variance de la population ( $\sigma$ ) était connue. En effet, caractériser la loi de  $\bar{X}$  en fonction de  $\sigma$  n'est intéressant que si l'on connaît  $\sigma$ . Continuons dans cette voie avant de se demander ce qui se passe si l'on ne connaît pas la valeur de la variance de la population.

Comme les tables de la loi normale ne donnent les valeurs que pour les lois centrées réduites, il faut se ramener à une loi centrée réduite. En utilisant les propriétés de la loi normale, on sait que si  $\bar{X}_n \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$  alors  $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0; 1)$ .

Rappel: pour obtenir une variable centrée réduite  $Z$  à partir d'une variable aléatoire  $X$ , on centre en soustrayant l'espérance (du coup, l'espérance de  $Z$  vaut 0) et on réduit en divisant par l'écart-type de  $X$  (du coup l'écart-type de  $Z$  vaut 1):

$$\text{Variable centrée réduite} = \frac{\text{v.a.r.-espérance de la v.a.r.}}{\text{E.T. de la v.a.r.}}$$

Le tableau 9 résume les résultats de distribution d'une moyenne empirique dans le cas où la variance de la population est connue, avec un tirage avec remise ou sans remise dans une population grande.

Table 9. Cas 1: la variance de la population est connue; tirage avec remise ou sans remise dans une population infinie

	Population normale	Autre population ( $n \geq 30$ )
Loi de proba de $\bar{X}$	Normale	Approx. normale
Espérance de $\bar{X}$		$E(\bar{X}) = \mu$
Variance de $\bar{X}$		$V(\bar{X}) = \sigma^2/n$
Fluctuations de l'écart réduit		$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0; 1)$

Même si cette formule reste vraie lorsque la variance est inconnue, elle n'est pas d'une grande utilité. On verra plus loin ce qu'on peut faire lorsque la variance de la population est inconnue (la formule change lorsqu'on remplace  $\sigma$  par une estimation de celle-ci).

N'oubliez pas que le cas sans remise pour une population de taille finie est connu, il suffit d'appliquer la correction pour la non-remise (voir plus haut).

## 2.5 Calculs de probabilités sur des moyennes d'échantillon

### 2.5.1 Événements associés à $\bar{X}$

Comme pour les variables aléatoires, les problèmes de probabilités basés sur des moyennes se ramènent à deux types de questions:

- on doit calculer la probabilité que  $\bar{X}$  soit supérieure ou inférieure à une certaine valeur numérique donnée;
- on doit calculer un seuil de probabilité tel qu'il y ait 90% (ou 95%, ou 99%) de chances d'obtenir une valeur qui soit supérieure ou inférieure au seuil.

Voir tableau des événements associés à  $\bar{X}$ .

### 2.5.2 Calcul du seuil de probabilité et application au test d'aptitudes.

Nous désignons par  $z_\alpha$  le seuil d'une loi normale centrée réduite tel que  $P(Z > z_\alpha) = \alpha$ .

Dans le cas d'une population normale ou d'un échantillon de taille supérieure à 30 et lorsqu'on connaît la variance, le seuil tel que la probabilité que  $\bar{X}$  soit supérieur à  $\alpha$  est:

$$\begin{aligned}k &= \mu + z_\alpha \cdot \sigma_{\bar{X}} \\ &= \mu + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\end{aligned}$$

Application: un spécialiste en psychologie industrielle utilise auprès d'une population d'ouvriers un test d'aptitudes à effectuer une certaine tâche. D'après les standards établis, les résultats aux tests sont distribués selon une loi normale d'espérance  $\mu = 150$  et de variance  $\sigma^2 = 100$ . On administre le test à un échantillon aléatoire de 25 ouvriers d'une entreprise. Le résultat moyen est de 154.7 avec un écart-type de 12.3.

- tableau comparatif des moments population/échantillon;
- caractéristiques de la moyenne d'échantillon;
- probabilité qu'un résultat varie entre 140 et 165;
- probabilité que le résultat moyen soit compris entre 145 et 154;
- seuil  $k$  tel que la probabilité d'obtenir une moyenne d'échantillon supérieure à  $k$  est 10%;
- probabilité d'obtenir une moyenne d'échantillon supérieure ou égale à 154.7;

### 3 Distribution d'échantillonnage d'une variance

#### 3.1 Définitions et exemple

**Définition 22** La variance empirique non modifiée de l'échantillon est:

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

et la variance empirique modifiée de l'échantillon est:

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

Attention: les notations pour la variance empirique et la variance empirique modifiée changent d'un livre à l'autre. On notera  $\tilde{s}^2$  la réalisation de la variance empirique et  $s^2$  celle de la variance empirique modifiée. Je vais aussi définir la **variance semi-empirique** (nom local, ce moment empirique n'est pratiquement pas utilisé) dans laquelle on suppose connue l'espérance de  $X$  (notée  $\mu$ ):

$$\mathcal{V}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Reprenons tout d'abord l'exemple des montants alloués aux individus pour un voyage scolaire. Nous pouvons de la même façon que nous avons calculé la moyenne sur divers échantillons, calculer la variance empirique et la variance empirique modifiée sur divers échantillons (voir tableau 10).

Table 10. Variances pour l'ensemble des échantillons

Echantillon n°	$(i, j, k)$	Résultat de l'échantillonnage	$\tilde{s}^2$	$s^2$
1	(1,2,3)	(600,150,300)	35000	52500
2	(1,2,4)	(600,150,600)	45000	67500
3	(1,2,5)	(600,150,150)	45000	67500
4	(1,3,4)	(600,300,600)	20000	30000
5	(1,3,5)	(600,300,150)	35000	52500
6	(1,4,5)	(600,600,150)	45000	67500
7	(2,3,4)	(150,300,600)	35000	52500
8	(2,3,5)	(150,300,150)	5000	7500
9	(2,4,5)	(150,600,150)	45000	67500
10	(3,4,5)	(300,600,150)	35000	52500

Bien entendu, à chaque fois, il suffit de multiplier la variance non modifiée par  $n/(n-1) = 3/2 = 1.5$  pour obtenir la variance modifiée. On verra dans un instant pourquoi on définit deux statistiques différentes pour la variance. On peut facilement calculer aussi les moments de ces variances échantillonnales. Il vient:

Table 11. Moments empiriques des variances empiriques

	Moyenne	Variance
$\tilde{S}^2$	34500	152250000
$S^2$	51750	342562500

Ainsi, de la même façon que la moyenne d'échantillon est une statistique et donc une variable aléatoire, la variance d'échantillon, modifiée ou non, est une variable aléatoire dont on peut calculer les moments et la loi. C'est l'objet des sections suivantes.

### 3.2 Espérance d'une variance échantillonnale

Dans cette section, vous allez comprendre pourquoi on définit une variance empirique et une variance modifiée. Nous allons, comme nous l'avons fait pour  $\bar{X}_n$ , calculer l'espérance de la variance. Mais commençons par regarder ce qui se passe pour la variance que j'ai appelée semi-empirique:

$$E(\mathcal{V}(X)) = E\left(\frac{1}{n} \sum_i (X_i - \mu)^2\right) = \frac{1}{n} \sum_i E((X_i - \mu)^2) = \frac{1}{n} \sum_i \sigma^2 = \sigma^2$$

L'espérance de la variance semi-empirique est  $\sigma^2$ . Cependant, on ne connaît généralement pas la valeur de  $\mu$ , ce qui nous conduit à la remplacer par son estimateur  $\bar{X}_n$ , d'où les variances empiriques modifiées et non modifiées. Ce sont les vraies variances empiriques et nous allons maintenant donner leur espérance:

**Proposition 23** *Soit  $X_i$  une population d'espérance  $\mu$  et de variance  $\sigma^2$ . La variance empirique calculée sur un échantillon aléatoire de taille  $n$  de cette population a pour espérance:*

$$E(\tilde{S}_n^2) = \frac{n-1}{n} \cdot \sigma^2.$$

**Preuve.** Soit  $\mu$  l'espérance du caractère  $X_i$ . On réécrit  $(X_i - \bar{X}_n)^2$  comme ceci:  $[(X_i - \mu) - (\bar{X}_n - \mu)]^2$ ; en développant l'expression, la variance empirique

s'écrit:

$$\begin{aligned}
 \tilde{S}_n^2 &= \frac{1}{n} \sum_i [(X_i - \mu)^2 - 2(X_i - \mu)(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2] \\
 &= \frac{1}{n} \sum_i (X_i - \mu)^2 - \frac{2}{n} (\bar{X}_n - \mu) \cdot \sum_i (X_i - \mu) + (\bar{X}_n - \mu)^2 \\
 &= \frac{1}{n} \sum_i (X_i - \mu)^2 - 2(\bar{X}_n - \mu) \cdot (\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2 \\
 &= \frac{1}{n} \sum_i (X_i - \mu)^2 - (\bar{X}_n - \mu)^2
 \end{aligned}$$

dont l'espérance vaut:

$$\begin{aligned}
 E(\tilde{S}_n^2) &= \frac{1}{n} \sum_i E((X_i - \mu)^2) - E((\bar{X}_n - \mu)^2) \\
 &= \frac{1}{n} \sum_i \sigma^2 - E((\bar{X}_n - \mu)^2) \\
 &= \sigma^2 - \frac{1}{n} \sigma^2 \\
 &= \frac{n-1}{n} \cdot \sigma^2
 \end{aligned}$$

en utilisant le résultat sur la variance de la moyenne. ■

Que remarque-t-on? que la variance de la moyenne ne vaut pas, en espérance, la variance de la population. On peut donc noter immédiatement que le remplacement de  $\mu$  par son estimation  $\bar{X}_n$  affecte l'espérance de la variance empirique, et ce même si  $E(\bar{X}_n) = \mu$ . C'est la raison pour laquelle on a défini la variance modifiée. En effet, nous avons vu que:

$$S_n^2 = \frac{n}{n-1} \cdot \tilde{S}_n^2$$

donc:

$$E(S_n^2) = \frac{n}{n-1} \cdot E(\tilde{S}_n^2) = \sigma^2$$

**Proposition 24** Soit  $X_i$  une population d'espérance  $\mu$  et de variance  $\sigma^2$ . La variance empirique modifiée calculée sur un échantillon aléatoire de taille  $n$  de cette population a pour espérance:

$$E(S_n^2) = \sigma^2.$$

La variance empirique modifiée, elle, vaut en espérance la variance de la population. Si on veut "deviner" la variance d'une population en utilisant un échantillon, il vaut donc mieux calculer la variance empirique modifiée! Notez

cependant que lorsque  $n$  devient grand, il n'y a pratiquement plus de différence entre les deux variances empiriques.

Les calculs pour obtenir la variance de la variance empirique et sa loi étant complexes, nous les omettons pour l'instant. Nous reviendrons dessus plus tard dans le cours.

## 4 Distribution d'échantillonnage d'une proportion

Jusqu'ici nous nous sommes concentrés sur la description d'échantillons à caractères quantitatifs. Lorsqu'on s'intéresse à la fréquence, c'est généralement que le caractère est qualitatif. Par exemple: vous voulez savoir quelle proportion de la classe de jeunes mannequins a les yeux bleus, est blonde, etc. Mais vous pouvez aussi disposer d'une information censurée. Si par exemple on imagine que cette classe est un échantillon de l'ensemble des MSG et que le caractère est "a plus de 10 en stats", on voit bien que l'on pourrait définir un autre caractère qui est la note exacte en stats. Cependant, si vous ne disposez pas de toute l'information, vous serez amenés à considérer le caractère "a la moyenne" pour estimer quelle proportion de la classe aura la moyenne. L'objet de la section est de décrire la distribution d'échantillonnage d'une nouvelle statistique qui est la **proportion d'échantillon**. Cette partie concerne à la fois la fréquence et la proportion puisque la proportion vaut la fréquence divisée par la taille de la population (ou échantillon).

### 4.1 Définition

**Définition 25** *On appelle proportion d'échantillon la statistique  $\hat{P}$  qui associe à chaque échantillon aléatoire de taille  $n$  prélevé dans une même population la proportion  $\hat{p}$  d'individus dans cet échantillon possédant un certain caractère qualitatif.*

Pourquoi ne s'intéresser à la proportion que dans le cas d'un caractère qualitatif? parce que lorsque l'on parle de caractère quantitatif, on a en tête une variable continue pour laquelle l'occurrence d'une certaine valeur est très improbable: quelle est la probabilité que vous mesuriez 1m65987 très exactement? probablement 0. La proportion et la fréquence ne sont adaptées pour décrire un échantillon que si celui-ci a des points d'accumulation. On pourrait par exemple se demander quelle proportion des salariés travaille au SMIC.

#### Remarques

- $\hat{P}$  est une variable aléatoire;
- $\hat{p}$  est une réalisation;

- $\hat{p} = \frac{x}{n}$  où  $x$  désigne le nombre d'individus dans l'échantillon de taille  $n$  possédant le caractère considéré.

## 4.2 La proportion est une moyenne

Commençons par montrer que la proportion empirique peut être vue comme la moyenne d'une certaine variable. Définissons  $X_i$  comme la variable suivante:

$$X_i = \begin{cases} 1 & \text{si l'individu } i \text{ possède le caractère} \\ 0 & \text{sinon} \end{cases}$$

Ce type de variables a tout un tas de noms dont: variable indicatrice et dummy. La fréquence d'un caractère dans un échantillon s'écrit alors

$$F_n = \sum_{i=1}^n X_i$$

et la proportion:

$$\hat{P} = \frac{1}{n} \sum_{i=1}^n X_i.$$

La proportion empirique est donc la moyenne de la variable  $X_i$ . À partir de là, il va être très facile de caractériser sa distribution au vu de ce qui a déjà été démontré dans le cours.

## 4.3 Espérance et variance d'une proportion empirique

Pour retrouver l'espérance et la variance de la proportion, on peut soit utiliser le fait que la proportion est une moyenne, soit utiliser le fait que la fréquence suit une binômiale.

### 4.3.1 La fréquence est une binômiale

Comme  $F_n$  suit une  $\mathcal{B}(n, p)$ , on sait que  $E(F_n) = np$  et  $V(F_n) = np(1-p)$ . Comme  $\hat{P} = \frac{1}{n}F$ , il vient:

$$\begin{aligned} E(\hat{P}) &= \frac{1}{n}E(F_n) = \frac{np}{n} = p \\ V(\hat{P}) &= \frac{1}{n^2}V(F_n) = \frac{p(1-p)}{n} \end{aligned}$$

et on a caractérisé les moments de  $\hat{P}$ .

### 4.3.2 La proportion est une moyenne

En utilisant les résultats obtenus précédemment dans le cours, on sait que:

$$E(\bar{X}) = E(X) \quad \text{et} \quad SE_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

comme on a montré que  $\hat{P} = \bar{X}$  pour  $X$  égal à la variable indicatrice de “possède le caractère”, on peut appliquer ces propriétés pour obtenir que si  $n$  est grand ou si le tirage se fait avec remise:

$$E(\hat{P}) = E(\bar{X}) = E(X) = p$$
$$V(\hat{P}) = V(\bar{X}) = \frac{V(X)}{n} = \frac{p(1-p)}{n}$$

en utilisant le fait que  $X$  suit une Bernoulli  $\mathcal{B}(p)$  et a donc pour espérance  $p$  et pour variance  $p(1-p)$ .

Pour résumer:

**Proposition 26** *Les moments de la proportion empirique sont:*

$$E(\hat{P}) = p$$
$$V(\hat{P}) = \frac{p(1-p)}{n}$$

## 4.4 Distribution de la proportion

En fait, les deux dernières approches indiquent implicitement les lois:

- la fréquence suit une binômiale donc  $n\hat{P} \rightsquigarrow \mathcal{B}(n, p)$
- lorsque  $n$  est grand, la proportion en tant que moyenne de  $X$ , suit une loi normale.

### 4.4.1 Approximation par une normale

**n grand?** Pour rappel, l'application du TCL nécessite d':

- avoir une population normalement distribuée et  $n$  quelconque
- soit avoir  $n$  grand et une distribution de population quelconque

On est dans le cas où la loi de  $X$  (Bernoulli) n'est pas normale, il faut donc que  $n$  soit grand. Ceci dit, on a vu aussi que plus la distribution de départ est symétrique et moins  $n$  doit être élevé pour s'approcher de façon raisonnable de la normale. Dans le cas d'une Bernoulli, la distribution n'est symétrique que si  $p = 0.5$ . Donc en pratique, plus  $p$  est loin de  $1/2$  et plus il faut que  $n$  soit élevé pour retenir l'approximation normale. En pratique, on retient que l'on peut appliquer la propriété lorsque  $np \geq 5$  et  $n(1-p) \geq 5$ .

Table 12. Approximation normale de la proportion; tirage avec remise ou sans remise dans une population infinie

Conditions	$p$ connue et $np \geq 5$ et $n(1-p) \geq 5$	$p$ inconnue et $np \geq 5$ et $n(1-p) \geq 5$
Distribution de $\hat{P}$	Approx. normale	Approx. normale
Espérance de $\hat{P}$	$E(\hat{P}) = p$	$E(\hat{P}) = p$
Variance de $\hat{P}$	$V(\hat{P}) = \frac{p(1-p)}{n}$	$V(\hat{P})$ estimée par $\frac{\hat{p}(1-\hat{p})}{n}$
Ecart réduit	$Z = \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0, 1)$	

**Correction pour continuité** Lorsque l'on utilise l'approximation par loi normale, il faut prendre en compte que l'on approxime une loi discrète par une loi continue et introduire le terme de correction pour continuité:

$$P(\hat{P}_n \leq x) = P\left(\frac{\hat{P}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{x - p + 0.5/n}{\sqrt{\frac{p(1-p)}{n}}}\right) \simeq F\left(\frac{x - p + 0.5/n}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

#### 4.4.2 Approximation par une Poisson

**Que faire si  $n$  est grand mais  $p$  petit?** Dans ce cas-la, il est difficile d'utiliser la loi binômiale qui ne sera pas tabulée pour  $n$  grand et vous êtes dans le cas où vous ne pouvez utiliser l'approximation normale. Vous pouvez alors utiliser l'approximation de la binômiale par la loi de Poisson.

Comme  $n\hat{P} \rightsquigarrow \mathcal{B}(n, p)$  alors  $n\hat{P} \rightsquigarrow_{app} \mathcal{P}(np)$ .