Chapitre 2. Échantillonnage

Pierre Duchesne

January 19, 2017

- En théorie des probabilités, une question intéressante pourrait être de la forme: *Si on lance une pièce de monnaie régulière à 1000 reprises, quelle est la probabilité d'observer 503 faces?*
- En statistique, la question intéressante serait plutôt: *Si on lance une pièce de monnaie à 1000 reprises et que l'on observe 503 faces, est-ce que la pièce est régulière?*

Les deux disciplines s'intéressent à un caractère étudié qui serait ici:

$$X = \text{nombre de faces en 1000 lancers de la pièce.}$$

Ainsi, en probabilités, on suppose que l'on connaît tous les paramètres qui définissent la loi de probabilité. Dans ce cas-ci, $X \sim Bin(1000, \frac{1}{2})$.
En statistique, on ne connaît pas tous les paramètres. On aurait ici $X \sim Bin(1000, p)$ avec $0 \leq p \leq 1$.

Typiquement, nous pouvons dire qu'il y a deux grands problèmes dans le domaine de la statistique:

1. **Théorie de l'estimation**: le problème du statisticien consiste à attribuer une valeur à un paramètre inconnu;
2. **Théorie des tests d'hypothèses**: le statisticien cherche à vérifier une conjecture sur le paramètre inconnu.

Les deux théories ci-dessus forment ce que l'on appelle **l'inférence statistique**.

- Lorsque l'on effectue de l'inférence statistique, nous devons disposer d'information sur le caractère à l'étude, et ceci s'obtient au moyen d'un échantillon.
- Dans l'exemple du lancer de la pièce, on pourrait avoir comme échantillon la chaîne décrivant les 1000 résultats: $PFPP \ldots F$, avec F pour face, et P pour pile.
- Alternativement, on pourrait poser:

$$X_i = \begin{cases} 1, & \text{si } F \text{ au tirage } i, \\ 0, & \text{sinon.} \end{cases}$$

- On aurait alors dans la chaîne de caractères précédente: $X_1 = 0$, $X_2 = 1$, $X_3 = X_4 = 0, \ldots, X_{1000} = 1$.

Chapitre 2.1 Échantillonnage dans une population infinie

Considérons une population *finie* pour laquelle on étudie le caractère X suivant:

$$X = \text{nombre d'enfants dans la famille.}$$

Dans notre exemple, la population $\mathcal{P}$ est composée de cinq familles.

unité	nb d'enfants
1	0
2	1
3	1
4	2
5	3

unité	nb d'enfants
1	0
2	1
3	1
4	2
5	3

Ainsi la loi de probabilité de X est:

unité	nb d'enfants
0	1/5
1	2/5
2	1/5
3	1/5

Supposons maintenant que l'on prélève un échantillon de taille $n = 2$ dans la population finie.
Deux plans d'échantillonnage sont possibles:

- Échantillonnage aléatoire simple *avec* remise;
- Échantillonnage aléatoire simple *sans* remise.

Le tirage aléatoire simple sans remise est sans doute plus naturel, mais le tirage avec remise offre des propriétés plus intéressantes (et surtout plus commodes mathématiquement).

Échantillonnage avec remise

Considérons X_1 et X_2 les variables aléatoires pour chacun des deux tirages.
Nous avons que X_1 a la même loi que la loi du caractère étudié X:

unité	nb d'enfants
0	1/5
1	2/5
2	1/5
3	1/5

De même, X_2 a la même loi que la loi du caractère étudié X.
De plus, X_1 et X_2 sont indépendantes.

Échantillonnage sans remise

Considérons encore X_1 et X_2 les variables aléatoires pour chacun des deux tirages.
Encore ici, X_1 est une variable aléatoire dont la loi est identique à celle de X.
Cependant, à prime abord, la loi de X_2 n'est plus aussi évidente.
En effet, dans un échantillonnage sans remise, la valeur que prendra X_2 dépend maintenant de la valeur qui aura été prise par X_1.

Afin de prendre la loi de X_2, une façon de procéder est de trouver la loi conjointe du couple (X_1, X_2), que l'on note $p_{X_1,X_2}(x_1, x_2)$.

$$p_{X_1,X_2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2|X_1 = x_1)$$

unité	nb d'enfants
1	0
2	1
3	1
4	2
5	3

On trouve:

$x_1 \setminus x_2$	0	1	2	3	$p_{X_1}(x_1)$
0	0/20	2/20	1/20	1/20	4/20 = 1/5
1	2/20	2/20	2/20	2/20	8/20 = 2/5
2	1/20	2/20	0/20	1/20	4/20 = 1/5
3	1/20	2/20	1/20	0	4/20 = 1/5
$p_{X_2}(x_2)$	1/5	2/5	1/5	1/5	20/20 = 1

On remarque que X_1 est de même loi que X, et il en va de même de X_2, qui a même loi que X.
Cependant, X_1 et X_2 sont dépendantes, et on peut montrer que:

$$\text{cov}(X_1, X_2) = -\frac{26}{100}.$$

Ainsi les tirages avec remise, quoique moins naturels, ont des propriétés mathématiques plus intéressantes.
Si la population était de taille infinie (en pratique si la population est très grande), tirer au hasard avec ou sans remise revient à toute fin pratique au même.

Échantillon aléatoire

On va donc considérer à partir de maintenant que l'on a affaire à des populations infinies.
Soit X un caractère étudié dans une population $\mathcal{P}$. Alors $\mathcal{E} : X_1, X_2, \ldots, X_n$ est un échantillon aléatoire de taille n si $X_1, X_2, \ldots, X_n$ sont indépendantes et chacune de même loi que X.
Un caractère X est une v.a. et on appellera $\mu_X = E(X)$ la moyenne de la population ou encore la moyenne théorique. De même, $\sigma_X^2 = E\{(X - \mu_X)^2\}$ est la variance de la population ou encore la variance théorique.

Moments théoriques

En général, si k est un entier, nous noterons:

$$\alpha_k = E(X^k) = \text{moment théorique d'ordre } k,$$
$$\mu_k = E\{(X - \mu_X)^k\} = \text{moment théorique centré d'ordre } k.$$

On note que $\alpha_1 = \mu_X$, et que $\mu_1 = 0$, et $\mu_2 = \sigma_X^2$.

Chapitre 2.2 Statistiques et moments expérimentaux

Une statistique est une variable aléatoire qui n'est fonction que d'un échantillon aléatoire $\mathcal{E} : X_1, X_2, \ldots, X_n$ dans une population $\mathcal{P}$.
Plus précisément, soit X est un caractère étudié dans une population $\mathcal{P}$. Si $\mathcal{E} : X_1, X_2, \ldots, X_n$ associé à X, alors une statistique $T = t(X_1, \ldots, X_n)$ où $t : \mathbb{R}^n \to \mathbb{R}$ est une fonction entièrement connue.

Exemples: Soit un échantillon aléatoire $\mathcal{E} : X_1, X_2, \ldots, X_n$ associé à X.

1. La fonction $t(x_1, \ldots, n_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ donne lieu à la statistique $T = \frac{1}{n}\sum_{i=1}^{n} X_i$, appelée la **moyenne échantillonnale**.
2. La fonction $t(x_1, \ldots, n_n) = \max_{1\leq i\leq n} x_i - \min_{1\leq i\leq n} x_i$. La statistique est $T = \max_{1\leq i\leq n} X_i - \min_{1\leq i\leq n} X_i$ et permet d'introduire **l'étendue**.
3. La fonction $t(x_1, \ldots, n_n) = \frac{1}{n}\sum x_i^k$ permet d'introduire les **moments expérimentaux** $T = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ (où k est un entier).
4. La fonction $t(x_1, \ldots, n_n) = \frac{1}{n}\sum (x_i - \mu_X)^k$ permet d'introduire les **moments expérimentaux centrés** $T = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu_X)^k$ (où k est un entier et μ_X est présumé connu).

Lorsque $k = 2$, la statistique $T = \frac{1}{n}\sum(X_i - \mu_X)^2$ est appelée la variance échantillonnale. On pose:

$$\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \mu_X)^2.$$

C'est une statistique en autant que μ_X est connue. Si μ_X est inconnue $\hat{\sigma}^2$ n'est plus une statistique et on introduit plutôt:

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2.$$

On appelle S^2 la variance échantillonnale corrigée.

Variance échantillonnale corrigée

On rappelle que:

$$\begin{aligned} E\{[X-\mu_X]^2\} &= E\{X^2-2\mu_X X+\mu_X^2\}, \\ &= E(X^2)-2\mu_X^2+\mu_X^2, \\ &= E(X^2)-\{E(X)\}^2. \end{aligned}$$

Les mêmes étapes donnent:

$$
\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\{X_i-\bar{X}\}^2, \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\{X_i^2-2\bar{X}X_i+\bar{X}^2\}, \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2\right\}, \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2\right\}, \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right\}, \\
&= \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}X_i^2-\bar{X}^2\right\}.
\end{aligned}
$$

En résumé: $\sigma^2 = E\{[X - \mu_X]^2\} = E(X^2) - \{E(X)\}^2$ est fonction du premier moment théorique et du second moment théorique (élevé au carré).
De même, $S^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right\}$ est fonction du premier moment échantillonnal et du second moment échantillonal (élevé au carré).

Espérance et variance de $\bar{X}$, $\hat{\sigma}^2$ et S^2

Soit un échantillon aléatoire $\mathcal{E} : X_1, X_2, \ldots, X_n$ associé à X.
Soit $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Alors:

$$\begin{aligned} E(\bar{X}) &= \mu_X, \\ \text{var}(\bar{X}) &= \frac{\sigma^2}{n}. \end{aligned}$$

Soit $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2$. Posons $\mu_k = E\{(X - \mu_X)^k\}$. Alors:

$$\begin{aligned} E(\hat{\sigma}^2) &= \sigma^2, \\ \text{var}(\hat{\sigma}^2) &= \frac{\mu_4 - \mu_2^2}{n}, \\ &= \frac{\mu_4 - \sigma^4}{n}. \end{aligned}$$

Espérance et variance de $\bar{X}$, $\hat{\sigma}^2$ et S^2 (suite)

Soit un échantillon aléatoire $\mathcal{E} : X_1, X_2, \ldots, X_n$ associé à X.
Soit $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Alors:

$$
\begin{aligned}
E(S^2) &= \sigma^2, \\
\text{var}(S^2) &= \left(\frac{n}{n-1} \right)^2 \left\{ \frac{\mu_4 - \mu_2^2}{n} - 2\left(\frac{\mu_4 - 2\mu_2^2}{n^2}\right) + \frac{\mu_4 - 3\mu_2^2}{n^3} \right\}
\end{aligned}
$$

Le résultat précédent sur la variance est général. Si on suppose que X est de loi normale, $X \sim \mathcal{N}(\mu_X, \sigma^2)$, alors

$$
\text{var}(S^2) = \frac{2\sigma^4}{n-1}.
$$