IUT HSE
Introduction aux probabilités et statistiques
Variables aléatoires

Philippe Jaming

Institut Mathématique de Bordeaux
Philippe.Jaming@gmail.com
http://www.math.u-bordeaux1.fr/~ pjaming/

Applications aux statistiques

X variable aléatoire de moyenne μ de variance σ.
But : obtenir des informations sur X à partir d'informations partielles : un échantillon.

Definition

$X_1, X_2, \dots, X_n$ n variables aléatoires. Un *échantillon* de taille n est une réalisation $(x_1, \dots, x_n)$ de ces variables.
La façon d'obtenir cet échantillon est appelé *échantillonage*.

Exemple 1

Une chaîne de fabrication produit 40 000 fours dont 2% sont défectueux, on note D un four défectueux, M un four qui fonctionne. Le service de contrôle-qualitée qui ne connaît pas ces chiffres, teste aléatoirement 100 fours pour estimer la qualitée de l'ensemble de la fabrication.Il obtient l'échantillon
$(M, M, D, M, M, M, M, D, M, M, \dots, M, D, M \dots)$.
X variable aléatoire : état d'un four.On a $\mathbb{P}[X = D] = 0.02$ et $\mathbb{P}[X = M] = 0.98$. Donc X suit une loi de Bernouilli $p = 0.02$
L'échantillon est la réalisation de $(X_1, \dots, X_{100})$ où chaque X_i est une copie indépendante de X.

Exemple 2

On considère une urne contenant des boules (numérotées, de couleurs distinctes,...).
On souhaite obtenir des informations sur sa contenance par échantillonnage.
Pour cela, on va piocher n boules dans l'urne.
X_1 sera la variable aléatoire contenant le résultat du premier tirage, X_2 celle contenant le résultat du deuxième tirage etc...
Le choix est aléatoire : pas plus de raison de choisir une boule qu'une autre (pas le cas, par ex. pour les sondages)
2 possibilités : soit on remet la boule après chaque tirage les X_i sont *tous de même loi et indépendantes* on dit que l'échantillonage est *non exhaustif*.
ou on ne remet pas la boule : l'échantillonage est *exhaustif*.

X variable aléatoire de moyenne μ de variance σ.
$X_1, X_2, \ldots, X_n$, n copies indépendates de X. (même loi, donc même moyenne et variance).
$x_1, x_2, \ldots, x_n$ un échantillon de taille n associé.

Definition

– la *moyenne empirique* de l'échantillon est la quantité

$$\tilde{\mu} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

– la *variance empirique* de l'échantillon est la quantité $\tilde{\sigma}^2$

$$\tilde{\sigma}^2 = \frac{(x_1 - \tilde{\mu})^2 + (x_2 - \tilde{\mu})^2 + \ldots + (x_n - \tilde{\mu})^2}{n}.$$

$$\tilde{\sigma}^2 = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - \tilde{\mu}^2.$$

Exemple

Soit une urne contenant 10 boules numérotées de 1 à 10. Soit X la variable aléatoire contenant le résultat d'un tirage. On considère un échantillon aléatoire nonexhaustif de taille 5 : (1, 4, 7, 9, 4). La moyenne empirique de l'échantillon est :

$$\tilde{\mu} = \frac{1 + 4 + 7 + 9 + 4}{5} = 5.$$

La variance empirique de l'échantillon est :

$$\tilde{\sigma}^2 = \frac{1^2 + 4^2 + 7^2 + 9^2 + 4^2}{5} - 5^2 = 7,6.$$

Rappelons que pour X, l'espérence est

$$\mu = \frac{1 + 2 + \cdots + 10}{2} = 5,5$$

et la variance

$$\sigma^2 = \frac{1^2 + 2^2 + \cdots + 10^2}{2} - (5.5)^2 = 8.25.$$

donc $\tilde{\mu} \simeq \mu$ et $\tilde{\sigma}^2 \simeq \sigma^2$.

Interprétation : $\tilde{\mu}$ est une réalisation de la variable aléatoire

$$\frac{X_1 + \cdots + X_n}{n}.$$

La variance empirique est la réalisation de la variable aléatoire

$$\frac{X_1^2 + \cdots + X_n^2}{n} - \tilde{\mu}^2.$$

R la variable aléatoire (discrète) telle que $\mathbb{P}[R = x_i] = 1/n$ (loi uniforme sur l'échantillon $x_1, \ldots, x_n$). Alors $\mathbb{E}[R] = \tilde{\mu}$ et $Var(R) = \tilde{\sigma}^2$.

Loi forte des grands nombres

Soit (X_n) une suite de variables aléatoires de même moyenne μ. Alors

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu$$

presque sûrement.

Interprétation : pour n assez grand, presque toute réalisation de $\frac{X_1+X_2+\cdots+X_n}{n} \simeq \mu$.
En particulier, $\tilde{\mu} \simeq \mu$ et $\tilde{\sigma}^2 \simeq \sigma^2$ quand n est assez grand.

Exercice

Soit X la variable aléatoire mesurant le temps d'attente avant d'être mis en relation avec un service après vente. On souhaite évaluer le temps d'attente moyen ainsi que la probabilité d'attendre cinq minutes ou plus. Pour cela on mesure aléatoirement 20 temps d'attente, on a donc un échantillon aléatoire non-exhaustif $(x_1, \ldots, x_{20})$ de taille 20 de la durée d'attente X :

i	1	2	3	4	5	6	7	8	9	10
x_i	6	2	18	1	8	9	2	14	13	6
i	11	12	13	14	15	16	17	18	19	20
x_i	0	15	4	14	2	5	5	28	5	0

Exercice

1. Donnez une estimation du temps d'attente moyen.
2. Donez une estimation de la probabilité d'attente supérieure à 5 minutes.
   Indication : introduire y_i qui vaut 1 si $x_i \geq 5$ et 0 sinon. Pourquoi la probabilité cherchée est-elle la moyenne de y.

Réponse : moyenne $7,85$ minutes, proba $0,8$.

Théorème central limite (TCL)

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même moyenne μ et même variance σ^2.
Soit $S_n = X_1 + X_2 + \cdots + X_n$ et $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$.
Alors $Z_n \to \mathcal{N}(0,1)$ en loi, c'est-à-dire, pour tous $a < b$,

$$\left| \mathbb{P}(Z_n \in [a,b]) - \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \right| \to 0$$

quand $n \to +\infty$.

Ceci dit que la moyenne de S_n est $n\mu$, sa variance $\sigma\sqrt{n}$ et, pour n assez grand, S_n se comporte comme la variable aléatoire gaussienne de moyenne $n\mu$ et de variance $\sigma\sqrt{n}$.
Dans la pratique, on l'utilise pour $n \geq 30$.

Application à l'échantillonnage : approche par intervalle de confiance
X une variable aléatoire d'espérance μ *inconnue* et de variance σ^2 connue ou nonet soit $(x_1, x_2, ..., x_n)$ un échantillon aléatoire non-exhaustif de taille n correspondant à des variables aléatoires indépendantes $X_1, X_2, \ldots, X_n$ de même loi que X.
On en déduit la moyenne empirique $\tilde{\mu} = \dfrac{x_1 + \cdots + x_n}{n}$.
On cherche maintenant un intervalle $I_\alpha = [\tilde{\mu} - \eta_\alpha, \tilde{\mu} + \eta_\alpha]$ tel que $\mathbb{P}[\mu \in I_\alpha] \geq 1 - \alpha$ (en général, $\alpha = 0.05$ voire 0.01).
on veut donc que $\mathbb{P}[|\mu - \tilde{\mu}| > \eta_\alpha] \leq \alpha$.

On introduit deux nouvelles variables aléatoires :

$$Y = \frac{X_1 + \cdots + X_n}{n}$$

$$Z = \frac{(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu)}{\sigma\sqrt{n}}.$$

$\tilde{\mu}$ est une réalisation de Y et (TCL) Z suit $\mathcal{N}(0, 1)$.
$\frac{\sigma Z}{\sqrt{n}} = Y - \tilde{\mu}$

$\tilde{\mu}$ est une réalisation de Y et $\mathbb{E}[Y] = \mu$ donc
$\mathbb{P}[|\mu - \tilde{\mu}| > \eta_\alpha] = \mathbb{P}[|Y - \tilde{\mu}| > \eta_\alpha] = \mathbb{P}\left[\left|\dfrac{\sigma Z}{\sqrt{n}}\right| > \eta_\alpha\right]$
$= \mathbb{P}\left[\left|Z > \dfrac{\sqrt{n}\eta_\alpha}{\sigma}\right|\right]$.
Mais Z suit $\mathcal{N}(0, 1)$ pour le quel on a des tables (de loi normale) et on lit u_α tel que $\mathbb{P}[|Z| > u_\alpha] = \alpha$. Par exemple
pour $\alpha = 0.01$, $u_\alpha = 2.5758$ et pour $\alpha = 0.05$, $u_\alpha = 1.96$
μ a 99% de chances de se trouver dans l'intervalle

$$\left[\tilde{\mu} - \frac{2.5758\sigma}{\sqrt{n}}, \tilde{\mu} + \frac{2.5758\sigma}{\sqrt{n}}\right].$$

Si σ n'est pas connu, on prend $\tilde{\sigma}$.

On ne cherche plus à obtenir des informations sur une seule variable aléatoire, mais

Objectif

Déterminer si deux variables aléatoires sont reliées

Exemples :

- taille et poids d'une population.
- niveau de revenu et espérence de vie.
- Moyenne au bac et durée totale des études...

On a 2 variables aléatoires X et Y : on prend un échantillon de la variable aléatoire (X, Y)

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_1 & \cdots & y_n \end{pmatrix}$$

donc $x_1, \ldots, x_n$ échantillon de X, $y_1, \ldots, y_n$ échantillon de Y.

Objectif

Décider s'il existe une relation linéaire entre X et Y : a, b tels que $Y = aX + b$.

— $\mu_X = \dfrac{x_1 + \ldots + x_n}{n}$ moyenne empirique de X
— $\sigma_X^2 = \frac{x_1^2 + \ldots + x_n^2}{n} - \mu_X^2$ variance empirique de X
— μ_Y moyenne empirique de Y, σ_Y variance empirique de Y.

Definition

– La *covariance empirique* de X et Y est donnée par

$$\sigma_{X,Y} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu_X)(y_j - \mu_Y).$$

– Le *coefficient de corrélation empirique* de X et Y est donnée par

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

Propriétés

- $-1 \leq \rho_{X,Y} \leq 1$;
- si $\rho_{X,Y} = \pm 1$ alors il existe a, b tels que $Y = aX + b$ (ou $X = aY + b$)
- **Conclusion :** si $|\rho_{X,Y}|$ est trop loin de 1 alors X et Y ne sont pas liés.
- $\mu_{aX+b} = a\mu_X + b$, $\sigma_{aX+b} = a\sigma_X \to \sigma_{aX+b,cY+d} = ac\sigma_{X,Y}$.
- $\to \rho_{aX+b,cY+d} = \frac{ac}{|ac|}\rho_{X,Y}$.

Lorsque $|\rho_{X,Y}|$ est assez proche de 1, on veut de plus trouver la relation linéaire.
On cherche à trouver a, b t.q. $D(a, b) = \sum_{j=1}^{n}(y_j - ax_j - b)^2$ soit minimal.

Théorème

Il existe un et un seul couple (a, b) tel que $D(a, b)$ soit minimal.La droite $y = ax + b$ est donnée par

1. *Elle passe par le point (μ_X, μ_Y)*
2. *Elle a pour pente $a = \dfrac{\sigma_{XY}}{\sigma_X^2} = \rho_{XY}\dfrac{\sigma_Y}{\sigma_X}$.*
3. *donc $b = \mu_Y - \mu_X\frac{\sigma_{XY}}{\sigma_X^2} = \mu_Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}\mu_X$.*

Definition

La droite $y = ax + b$ est appelée *droite de régression* de l'échantillon $(x_1, \ldots, x_n)$, $(y_1, \ldots, y_n)$.

Exemple

On cherche à savoir s'il y a une corrélation linéaire entre le nombre de machines à laver et le nombre de déficients visuels dans une population. Pour cela, on a l'échantillon suivant relevé dans un pays d'Europe :

Année	1970	71	72	73	74	75	76
Machines (en millier)	13	20	23	25	27	31	36
Déficients (pour 1000 hab)	8	8	9	10	11	11	12
Année	77	78	79	1980	81	82	83
Machines (en millier)	46	55	63	70	76	81	85
Déficients (pour 1000 hab)	16	18	19	20	21	22	23

Calculons le coefficient de corrélation de ces échantillons.

X le nombre de machines, Y la proportion de déficients visuels.

$\mu_X = \frac{13+20+23+25+27+31+36+46+55+63+70+76+81+85}{14} = 46,5$

$\mu_Y = \frac{8+8+9+10+11+11+12+16+18+19+20+21+22+23}{14} = 14.86$

$x_i - \mu_X$	-33,5	-26,5	-23,5	-21.5	-19.5	-15.5	-10.5
$y_i - \mu_Y$	-6.86	-6.86	-5.86	-4.86	-3.86	-3.86	-2.86
$(x_i - \mu_X)(y_i - \mu_Y)$	230	182	138	104	75.2	59.8	30
$x_i - \mu_X$	-0.5	8.5	16.5	23.5	29.5	34.5	38.5
$y_i - \mu_Y$	1.14	3.14	4.14	5.14	6.14	7.14	8.14
$(x_i - \mu_X)(y_i - \mu_Y)$	-0.57	26.7	68.4	121	181	246	314

$\sigma_X^2 = \frac{1}{14}\sum(x_i - \mu_X)^2 = 616.12$

$\sigma_Y^2 = \frac{1}{14}\sum(y_i - \mu_Y)^2 = 30.75$

$\sigma_{X,Y} = \frac{1}{14}\sum(x_j - \mu_X)(y_j - \mu_Y) = 126.76$

$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = 0.92.$