9. Distributions d'échantillonnage

MTH2302D

S. Le Digabel, École Polytechnique de Montréal

A2017

(v3)

Plan

Introduction

But

Tirer des conclusions au sujet d'une population sans avoir à examiner toutes les unités expérimentales (difficile ou impossible).

Comment ?

On prélève un sous-ensemble (échantillon) de la population et on tire des conclusions sur la population à partir des résultats obtenus avec l'échantillon.

Par exemple, on estime la moyenne de la population avec la moyenne échantillonnale.

Définition d'un échantillon aléatoire

Un *échantillon aléatoire* de taille n de la variable aléatoire X est une suite de variables aléatoires indépendantes $X_1, X_2, \ldots, X_n$ ayant toutes la même distribution que X.

Une suite $x_1, x_2, \ldots, x_n$ de valeurs prises par les v.a. X_i est une *réalisation* de l'échantillon.

Remarque

On suppose habituellement que la population est infinie ou que la taille de l'échantillon est beaucoup plus petite que la taille de la population.

Exemple 1

On fait l'hypothèse que la taille (en cm) des 4000 étudiants masculins d'une école de génie est une variable aléatoire X distribuée normalement, c'est-à-dire que $X \sim \mathsf{N}(\mu, \sigma^2)$.

Un échantillon aléatoire de taille 50 de cette population est une suite de 50 variables aléatoires $X_i \sim \mathsf{N}(\mu, \sigma^2)$, $i = 1, 2, \ldots, 50$.

Paramètres d'une population

- Une population (v.a.) est *connue* si on connaît sa distribution, c'est-à-dire sa fonction de masse ou de densité.
- En pratique on peut connaître une population seulement partiellement, c'est-à-dire qu'on connaît la forme générale de sa distribution mais avec des *paramètres* inconnus.

Exemple 2

On fait l'hypothèse que la taille des étudiant est distribuée normalement : $X \sim \mathrm{N}(\mu, \sigma^2)$ mais on ne connaît pas les paramètres μ et σ^2 (moyenne et variance).

Ce sont ces paramètres que l'on cherche à estimer.

Définition d'une statistique

Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire d'une variable aléatoire X.

Une *statistique* est une fonction $h(X_1, X_2, \ldots, X_n)$ ne dépendant que des variables aléatoires X_i.

Exemples de statistiques :

- La moyenne échantillonnale $\overline{X} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i$
- La variance échantillonnale $S^2 = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (X_i - \overline{X})^2$
- La médiane échantillonnale, etc.

Distribution échantillonnale

Notion importante

Puisque les X_i sont des variables aléatoires, toute statistique est aussi une variable aléatoire et on s'intéresse à sa distribution, appelée *distribution échantillonnale*.

Par exemple, on discute dans les prochaines sections de l'espérance et la variance de la moyenne et la variance échantillonnales, c'est à dire $\mathsf{E}(\overline{X})$, $\mathsf{V}(\overline{X})$, $\mathsf{E}(S^2)$, et $\mathsf{V}(S^2)$.

Distribution échantillonnale de la moyenne $\overline{X}$

Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire d'une v.a. X de moyenne $\mu = \mathsf{E}(X)$ et variance $\sigma^2 = \mathsf{V}(X)$.

Soit $\overline{X}$ la moyenne échantillonnale. Alors

1. $\mathsf{E}(\overline{X}) = \mu$ ($\overline{X}$ *est un estimateur non-biaisé de* μ)
2. $\mathsf{V}(\overline{X}) = \dfrac{\sigma^2}{n}$

Ces résultats découlent directement des règles de combinaisons linéaires.

Exemple 3

Une population est constituée des nombres 2, 3, 6, 8, 11.

L'ensemble des échantillons (avec remise) de taille 2 est

(2,2)	(3,2)	(6,2)	(8,2)	(11,2)
(2,3)	(3,3)	(6,3)	(8,3)	(11,3)
(2,6)	(3,6)	(6,6)	(8,6)	(11,6)
(2,8)	(3,8)	(6,8)	(8,8)	(11,8)
(2,11)	(3,11)	(6,11)	(8,11)	(11,11) .

Calculer

1. La moyenne et la variance de la population : μ et σ^2.
2. L'espérance et la variance de la moyenne échantillonnale $\overline{X}$: $\mathrm{E}(\overline{X})$ et $\mathrm{V}(\overline{X})$.

Distribution de la moyenne $\overline{X}$ (suite)

En utilisant le théorème central limite, on peut donner la loi de probabilité de la moyenne échantillonnalle.

Si l'échantillon est suffisamment grand, $\overline{X}$ suit approximativement une loi $\mathrm{N}(\mu, \sigma^2/n)$.

Remarques

- On a aussi (approx.) $n\overline{X} \sim \mathrm{N}(n\mu, n\sigma^2)$.
- Si $X \sim N(\mu, \sigma^2)$, alors $\overline{X}$, et $n\overline{X}$ sont exactement normales, même pour de petits échantillons.

Distribution de la moyenne $\overline{X}$ (suite)

On peut également définir la variable aléatoire

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

qui suit approximativement une loi N$(0, 1)$.

Remarques

- Si $X \sim N(\mu, \sigma^2)$, alors Z est exactement normale, même pour de petits échantillons.
- On appelle *pivot* une variable aléatoire qui se calcule à partir d'une statistique et des paramètres de la population.
- Nous verrons qu'un pivot dont la loi de probabilité ne dépend pas des paramètres de la population permet de définir un *intervalle de confiance*.

Exemple 4

Toujours avec $X \sim \mathsf{N}(\mu, \sigma^2)$, supposons que l'on connaisse la moyenne et la variance de la population : $\mu = 175$ et $\sigma^2 = 10^2$.

On choisit 10 échantillons aléatoires de 50 étudiants chacun.

Pour combien de ces échantillons s'attend-on à avoir une moyenne comprise entre 174 et 176 cm ?

Distribution échantillonnale de la variance S^2

Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire d'une v.a. X de moyenne $\mu = \mathsf{E}(X)$, de variance $\sigma^2 = \mathsf{V}(X)$ et de coefficient d'aplatissement $\beta_2 = \frac{\mu_4}{\sigma^4}$.

Soit S^2 la variance échantillonnale. Alors

1. $\mathsf{E}(S^2) = \sigma^2$ (S^2 *est un estimateur non-biaisé de* σ^2)
2. $\mathsf{V}(S^2) = \sigma^4 \left(\dfrac{2}{n-1} + \dfrac{\beta_2 - 3}{n} \right)$

Remarques

- On peut montrer (difficile !) que S^2 suit approximativement une loi normale pour de grands échantillons.
- En supposant que X suit une loi normale, on peut définir la distribution de S^2 pour de petits échantillons.

Exemple 5

Une population est constituée des nombres 2, 3, 6, 8, 11.

Les variances échantillonnales

$$S^2 = \frac{1}{2-1}\left((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2\right)$$

des 25 échantillons (avec remise) de taille 2 sont :

0	0.5	8	18	40.5
0.5	0	4.5	12.5	32
8	4.5	0	2	12.5
18	12.5	2	0	4.5
40.5	32	12.5	4.5	0 .

Retrouver manuellement ces valeurs et calculer $\mathsf{E}(S^2)$.

La fonction gamma

Rappel

La *fonction gamma* est définie pour tout $x > 0$ par

$$\Gamma(x) = \int_{t=0}^{\infty} t^{x-1} e^{-t}\, dt.$$

Propriétés

1. $\Gamma(1) = 1$,
2. $\Gamma(1/2) = \sqrt{\pi}$,
3. $\Gamma(x) = (x-1)\Gamma(x-1)$ pour $x > 1$,
4. Si $x = n \in \mathbb{N}$ alors $\Gamma(n) = (n-1)!$,
5. Voir page 139 (2ème édition) / page 142 (3ème édition).

La loi du khi-deux

Soit $Z_1, Z_2, \ldots, Z_k$ des variables aléatoires indépendantes et identiquement distribuées selon une loi normale $\mathsf{N}(0,1)$. Alors la variable aléatoire

$$W = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

suit une *loi du khi-deux à k degrés de liberté*. On note $W \sim \chi_k^2$. La fonction de densité de W est

$$f(w) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma(k/2)} w^{(k/2)-1} e^{-w/2} & \text{si } w \geq 0 \text{ ,} \\ 0 & \text{sinon.} \end{cases}$$

Remarques : $\chi_1^2 \equiv (\mathsf{N}(0,1))^2$ et $\chi_k^2 \equiv \Gamma(k/2, 1/2)$. De plus, si $k/2$ est entier, alors $X_1 + X_2 + \ldots + X_{k/2} \sim \chi_k^2$ avec $X_i \sim \mathsf{Exp}(1/2)$, $i \in \{1, 2, \ldots, n\}$.

La loi du khi-deux (suite)

Soit $W \sim \chi^2_k$. Alors

1. $\mathsf{E}(W) = k$.
2. $\mathsf{V}(W) = 2k$.
3. Le *quantile* $\chi^2_{\alpha;k}$ est défini par $P(W > \chi^2_{\alpha;k}) = \alpha$ avec $0 \leq \alpha \leq 1$.

Calculs avec la loi du khi-deux

- Le tableau de la page 478 (2ème édition) / page 514 (3ème édition) donne les quantile de la loi du khi-deux.
- En R : $\chi^2_{\alpha;k}$ est donné par qchisq(1-α,k).
- En Excel : LOI.KHIDEUX.INVERSE.DROITE(α,k).

Exemple 6

Calculer $\chi^2_{0.1;3}$ et $P(X \leq 11.07)$ si $X \sim \chi^2_5$.

Additivité la loi du khi-deux

Théorème

Soient $W_1, W_2, \ldots, W_p$ des v.a. khi-deux à $k_1, k_2, \ldots, k_p$ degrés de liberté respectivement. Alors

$$Y = W_1 + W_2 + \ldots + W_p$$

suit une loi du khi-deux à $k = k_1 + k_2 + \ldots + k_p$ degrés de liberté.

Additivité la loi du khi-deux

Application du théorème d'additivité

Soit $Z_1, Z_2, \ldots, Z_n$ un échantillon aléatoire de $Z \sim N(0,1)$.
On définit

$$A = \sum_{i=1}^{n} Z_i^2 \ , \ B = \sum_{i=1}^{n} (Z_i - \overline{Z})^2 \text{ et } C = n(\overline{Z})^2.$$

On peut montrer que $A = B + C$.

De plus, $A \sim \chi_n^2$ et $C \sim \chi_1^2$.

On en déduit, d'après le théorème précédent, que $B \sim \chi_{n-1}^2$, car seule la loi χ_{n-1}^2, additionnée à une loi χ_1^2, peut donner une loi χ_n^2.

Distribution de la variance S^2 (suite)

Théorème

Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire de taille n d'une variable aléatoire normale $X \sim \mathrm{N}(\mu, \sigma^2)$ et S^2 la variance échantillonnale. Alors la variable aléatoire

$$W = \frac{(n-1)S^2}{\sigma^2}$$

suit une loi khi-deux avec $n-1$ degrés de liberté.

Distribution de la variance S^2 (suite)

Le théorème précédent nous permet de caractériser la distribution échantillonnale de S^2.

Soit $W \sim \chi^2_{n-1}$, avec $\mathsf{E}(W) = n-1$ et $\mathsf{V}(W) = 2(n-1)$. On a :

- $\mathsf{P}(S^2 \leq b) = \mathsf{P}\left(\frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b}{\sigma^2}\right) = \mathsf{P}\left(W \leq \frac{(n-1)b}{\sigma^2}\right)$
- $\mathsf{E}(S^2) = \mathsf{E}\left(\frac{\sigma^2}{n-1}W\right) = \frac{\sigma^2}{n-1}\mathsf{E}(W) = \sigma^2$
- $\mathsf{V}(S^2) = \mathsf{V}\left(\frac{\sigma^2}{n-1}W\right) = \frac{\sigma^4}{(n-1)^2}\mathsf{V}(W) = \frac{2\sigma^4}{n-1}$

Remarque : Ces résultats ne sont valides que si la population X suit une loi $\mathsf{N}(\mu, \sigma^2)$.

Exemple 7

On fait l'hypothèse que la taille (en cm) des 4000 étudiants masculins d'une école de génie est une variable aléatoire normale X de moyenne 175 et variance 10^2, c'est-à-dire $X \sim \mathsf{N}(\mu = 175, \sigma^2 = 10^2)$.

On choisit 10 échantillons de taille 50 de la population X.

Pour combien de ces échantillons s'attend-on à avoir une variance échantillonnale S^2 d'au plus 101 ?

Loi t de Student

Rappel

Si $X_1, X_2, \ldots, X_n$ est un échantillon aléatoire de taille n de la variable aléatoire X, où $\mathsf{E}(X) = \mu$ et $\mathsf{V}(X) = \sigma^2$, alors

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

suit approximativement une loi $\mathsf{N}(0, 1)$. Cette variable aléatoire est un pivot permettant de définir un *intervalle de confiance* pour μ.

Loi t de Student (suite)

Si la variance σ^2 de la population n'est pas connue, on remplace σ par l'écart-type échantillonal $S = \sqrt{S^2}$, S^2 étant la variance échantillonnale.

On obtient alors la variable aléatoire

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Cette v.a. est approximativement normale si n est suffisamment grand. Si $X \sim N(\mu, \sigma^2)$, on peut montrer que T suit une loi de Student. Cette loi est valide pour les petits et les grands échantillons.

Loi t de Student (suite)

Soit Z une variable aléatoire normale $\mathsf{N}(0,1)$ et W une variable aléatoire khi-deux à k degrés de liberté. Si Z et W sont indépendantes alors la variable aléatoire

$$T = \frac{Z}{\sqrt{W/k}}$$

suit une *loi t de Student avec k degrés de liberté*. On note $T \sim t_k$.

La fonction de densité de T est

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(\frac{t^2}{k}+1\right)^{-\frac{k+1}{2}}$$

pour tout $t \in \mathbb{R}$.

Loi t de Student (suite)

Soit $T \sim t_k$. Alors

1. $\mathsf{E}(T) = 0$.
2. $\mathsf{V}(T) = \frac{k}{k-2}$ pour $k > 2$ (variance infinie pour $k = 1$ et 2).
3. On définit le *quantile* $t_{\alpha;k}$ de T par $P(T > t_{\alpha;k}) = \alpha$ avec $0 \leq \alpha \leq 1$.

Propriété

La fonction de densité $f(t)$ est symétrique par rapport à sa moyenne 0 et alors $-t_{\alpha;k} = t_{1-\alpha;k}$.

Théorème

La loi t_k est approximativement identique à une loi normale $\mathsf{N}(0, 1)$ lorsque k est grand.

Calculs avec la loi de Student

Si on cherche le quantile $t_{\alpha;k}$ tel que $P(t_k > t_{\alpha;k}) = \alpha$:

- Le tableau à la page 479 (2ème édition) / page 515 (3ème édition) donne les quantiles $t_{\alpha;k}$.
- En R : $t_{\alpha;k}$ est donné par qt(1-α,k).
- En Excel : -LOI.STUDENT.INVERSE.N(α,k).

Exemple 8

Calculer $t_{0.9;3}$ et $P(X \leq 2.015)$ si $X \sim t_5$.

Utilisation de la loi de Student

Théorème

Soit $X_1, X_2, \ldots, X_n$ un échantillon de taille n d'une variable aléatoire normale $X \sim \mathsf{N}(\mu, \sigma^2)$. Soit aussi $\overline{X}$ et S^2 la moyenne et la variance échantillonnale. On peut montrer que $\overline{X}$ et S^2 sont indépendantes, de sorte que la statistique

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

suit une loi de Student avec $n-1$ degrés de liberté.

Exemple 9

Supposons que l'on s'intéresse maintenant à la taille (en cm) des 2000 étudiantes d'une école de génie.

On suppose que la taille X suit une loi normale de moyenne $\mu = 170$. La variance est inconnue.

Si on choisit un échantillon de taille 25 de cette population, quelle est la probabilité que le rapport

$$\frac{\overline{X} - 170}{S}$$

soit inférieur à 0.26 ?

Distribution d'une différence de moyennes

Considérons maintenant deux échantillons aléatoires indépendants $X_1, X_2, \ldots, X_{n_X}$ et $Y_1, Y_2, \ldots, Y_{n_Y}$ de deux variables aléatoires X et Y de moyenne et variance μ_X, σ_X^2 et μ_Y, σ_Y^2 respectivement.

On s'intéresse à la différence des moyennes échantillonnales $\overline{X} - \overline{Y}$.

Théorème

Dans la situation décrite ci-dessus

1. $\mathsf{E}(\overline{X} - \overline{Y}) = \mu_X - \mu_Y$.
2. $\mathsf{V}(\overline{X} - \overline{Y}) = \dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}$.

Distribution d'une différence de moyennes (suite)

Théorème

La variable aléatoire

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

suit approximativement une loi normale $\mathsf{N}(0,1)$ lorsque n_X et n_Y sont grands.

Remarque : Z suit exactement une loi $\mathsf{N}(0,1)$ si $X \sim \mathsf{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathsf{N}(\mu_Y, \sigma_X^2)$.

Exemple 10

Soit $X \sim \mathsf{N}(175, 10^2)$ et $Y \sim \mathsf{N}(170, 9^2)$ la taille (en cm) des étudiants et étudiantes d'une école de génie.

On choisit un échantillon de taille 50 de X et un échantillon de taille 25 de Y.

Quelle est la probabilité que la différence $\overline{X} - \overline{Y}$ soit inférieure à 4 cm ?

Distribution d'un rapport de variances

Considérons à nouveau deux échantillons aléatoires indépendants, de taille n_X et n_Y, des variables aléatoires X et Y.

On suppose que X et Y suivent des lois normales $\mathsf{N}(\mu_X, \sigma_X^2)$ et $\mathsf{N}(\mu_Y, \sigma_Y^2)$ respectivement.

On s'intéresse au rapport des variances échantillonnales S_X^2/S_Y^2.

Loi de Fisher

Soient U et V deux variables aléatoires indépendantes suivant une loi du khi-deux avec u et v degrés de liberté, respectivement. Alors la variable aléatoire

$$Y = \frac{U/u}{V/v}$$

suit une *loi de Fisher* à u et v degrés de liberté. On note $Y \sim F_{u,v}$. La fonction de densité Y est

$$f(y) = \begin{cases} \dfrac{\Gamma(\frac{u+v}{2})\left(\frac{u}{v}\right)^{u/2}}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})} y^{(u/2)-1}\left(\left(\dfrac{u}{v}\right)y+1\right)^{-(u+v)/2} & \text{si } y \geq 0 \,, \\ 0 & \text{si } y < 0. \end{cases}$$

Loi de Fisher (suite)

Soit $Y \sim F_{u,v}$. Alors

1. $\mathsf{E}(Y) = \dfrac{v}{v-2}$ si $v > 2$.
2. $\mathsf{V}(Y) = \dfrac{2v^2(u+v-2)}{u(v-2)^2(v-4)}$ si $v > 4$.
3. Le *quantile* $F_{\alpha;u,v}$ est défini par $P(Y > F_{\alpha;u,v}) = \alpha$ avec $0 \leq \alpha \leq 1$.

Propriété

Par la définition de la loi de Fisher, $1/Y \sim F_{v,u}$ et on trouve que

$$F_{1-\alpha;u,v} = \frac{1}{F_{\alpha;v,u}} \quad \text{(attention à l'inversion des indices !)}$$

Calculs avec la loi de Fisher

Si on cherche le quantile $F_{\alpha;u,v}$ tel que $P(Y > F_{\alpha;u,v}) = \alpha$:

- Les quantiles de $F_{u,v}$ sont donnés à la page 480 (2ème édition) / page 516 (3ème édition).
- En R : $F_{\alpha;u,v}$ est donné par qf(1-α,u,v).
- En Excel : $F_{\alpha;u,v}$ est donné par INVERSE.LOI.F.N(1-α,u,v).

Exemple 11

Calculer $F_{0.75;11,10}$ et $P(X \leq 200)$ si $X \sim F_{2,1}$.

Distribution d'un rapport de variances (suite)

Théorème

Soit $X_1, X_2, \ldots, X_{n_X}$ et $Y_1, Y_2, \ldots, Y_{n_Y}$ deux échantillons aléatoires indépendants, de taille n_X et n_Y, des variables aléatoires X et Y.

On suppose que X et Y suivent des lois normales $\mathrm{N}(\mu_X, \sigma_X^2)$ et $\mathrm{N}(\mu_Y, \sigma_Y^2)$ respectivement.

Soit S_X^2 et S_Y^2 les variances échantillonnales. Alors la variable aléatoire

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

suit une loi de Fisher à $n_X - 1$ et $n_Y - 1$ degrés de liberté.

Exemple 12

Soit $X \sim \mathsf{N}(175, 10^2)$ et $Y \sim \mathsf{N}(170, 9^2)$ la taille (en cm) des étudiants et étudiantes d'une école de génie.

On choisit un échantillon de taille 50 de X et un échantillon de taille 25 de Y.

Quelle est la probabilité que le rapport $\dfrac{S_X^2}{S_Y^2}$ soit inférieur à 3 ?