Biostatistique

Inférence statistique: distributions d'échantillonnage

Anicet Ebou, Institut National Polytechnique Félix Houphouët-Boigny,
ediman.ebou@inphb.ci

Plan

Statistique inférentielle 3
Échantillons aléatoires 5
Statistiques et distributions échantillonnales 8
Distribution échantillonnale de la moyenne 11
Distribution échantillonnale de la variance 16
Loi t de Student 25
Distribution échantillonnale d'une différence de deux moyennes 32
Distribution échantillonnale d'un rapport de variances 36

Statistique inférentielle

Présentation

But: Tirer des conclusion au sujet d'une population sans avoir à examiner toute la population.

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.

Exemple: On estime la moyenne de la population avec la moyenne échantillonnale.

Échantillons aléatoires

Définition

Definition 1: Echantillon aléatoire.

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

Une suite $x_1, x_2, ..., x_n$ 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: On fait l'hypothèse que la taille (en cm) de 4000 habitants d'une bourgarde est une variable aléatoire X distribuée normalement, c'est-à-dire que $X \sim N(\mu, \sigma^2)$. Un échantillon aléatoire de taille 50 de cette population est une suite de 50 variables aléatoires $X_i \sim N(\mu, \sigma^2), i = 1, 2, ..., 50$.

Paramètres d'une population

- Une population (variable aléatoire) 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: Dans l'exemple précédent, on a fait l'hypothèse que la taille des habitants est distribuée normalement: $X \sim N(\mu, \sigma^2)$ mais on ne connaît pas les paramètres μ et σ^2 (moyenne et variance).
Ce sont ces paramètres que l'on cherche à estimer.

Statistiques et distributions échantillonnales

Définition d'une statistique

Definition 2: Une statistique.

Soit $X_1, X_2, ..., X_n$ un échantillon aléatoire d'une variable aléatoire X.
Une *statistique* est une fonction $h(X_1, X_2, ..., X_n)$ ne dépendant que des variables aléatoires X_i.

Exemples de statistiques:

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

Distribution échantillonnale

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 $\mathbb{E}\left(\overline{X}\right), \mathbb{V}\left(\overline{X}\right), \mathbb{E}(S^2)$, et $\mathbb{V}(S^2)$.

Distribution échantillonnale de la moyenne

Distribution échantillonnale de la moyenne

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

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

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

Ces résultats découlent directement des règles de combinaisons linéaires.

Distribution échantillonnale de la moyenne (ii)

Exercice 1: Une population est constituée des nombres 2, 3, 6, 8 et 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 $\overline{X}$: $\mathbb{E}\left(\overline{X}\right)$ et $\mathbb{V}\left(\overline{X}\right)$.

Distribution échantillonnale de la moyenne (iii)

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 $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Remarques:

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

On peut également définir la variable aléatoire

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

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

Distribution échantillonnale de la moyenne (iv)

Remarques:

1. Si $X \sim N(\mu, \sigma^2)$, alors Z est exactement normale, même pour de petits échantillons.
2. On appelle *pivot* une variable aléatoire qui se calcule à partir d'une statistique et des paramètres de la population.
3. 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*.

Exercice 2:

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

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

Distribution échantillonnale de la variance

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

Soit S^2 la variance échantillonnale. Alors

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

Remarques:

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

Distribution échantillonnale de la variance (ii)

Exercice 3: Une population est constituée des nombres 2, 3, 6, 8 et 11. Les variances échantilonnales

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

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 $\mathbb{E}(S^2)$.

La fonction gamma (Γ)

Definition 3: Fonction gamma.

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

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

Propriétés:

1. $\Gamma(1) = 1$
2. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
3. $\Gamma(x) = (x-1)\Gamma(x-1)$ pour $x > 1$
4. Si $x = n \in \mathbb{N}$ alors $\Gamma(n) = (n-1)!$

Loi du khi-deux

Definition 4: Loi du khi-deux.

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

$$W = Z_1^2 + Z_2^2 + ... + Z_k^2$$

suit une *loi du khi-deux à k degrés de liberté*, noté $W \sim \chi_k^2$.

La fonction de densité de W est

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

Remarques: $\chi_1^2 \equiv (N(0,1))^2$ et $\chi_k^2 \equiv \Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$. De plus, si $\frac{k}{2}$ est entier, alors $X_1 + X_2 + ... + X_{\frac{k}{2}} \sim \chi_k^2$ avec $\chi_i \sim \exp\left(\frac{1}{2}\right)$, $i \in \{1, 2, ..., n\}$.

Loi du khi-deux (ii)

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

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

Calculs avec la loi du khi-deux:

- Les quantiles de la loi du khi-deux sont données en annexe de ce cours sous la forme d'une table.
- En utilisant le logiciel R: $\chi^2_{\alpha;k}$ est donnée par `qchisq(1-α, k)`.
- En utilisant le logiciel Excel: `LOI.KHIDEUX.INVERSE.DROITE(α, k)`.

Exemple: Calculer $\chi^2_{0,1;3}$ et $P(X \leq 11,07)$ si $X \sim X^2_5$.

Additivité de la loi du khi-deux

Theorem 1.

Soient $W_1, W_2, ..., W_p$ des variables aléatoires khi-deux à $k_1, k_2, ..., k_p$ degrés de liberté respectivement. Alors

$$Y = W_1 + W_2 + ... + W_p$$

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

Application du théorème d'additivité:
Soit $Z_1, Z_2, ..., Z_n$ un échantillon aléatoire de $Z \sim N(0,1)$. On définit

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

On peut montrer que $A = B + C$. De plus, $A \sim \chi^2_n$ et $C \sim \chi^2_1$. On en déduit, d'après le théorème précédent, que $B \sim \chi^2_{n-1}$, car seule la loi χ^2_{n-1}, additionnée à une loi χ^2_1, peut donner une loi X^2_n.

Distribution de la variance S^2 (suite)

Theorem 2.

Soit $X_1, X_2, ..., X_n$ un échantillon aléatoire de taille n d'une variable aléatoire normale $X \sim N(\mu, \sigma^2)$ 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.

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

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

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

Distribution de la variance S^2 (suite) (ii)

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

Exercice 4:
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 $X \sim N(\mu = 175, \sigma^2 = 10)$.
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

Loi t de Student

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

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

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

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.

Loi t de Student (ii)

On obtient alors la variable aléatoire

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

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

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

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

Loi t de Student (iii)

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 $t \in \mathbb{R}$.
Soit $T \sim t_k$. Alors

1. $\mathbb{E}(T) = 0$
2. $\mathbb{V}(T) = \frac{k}{k-2}$ pour k > 2 (variance infinie pour $k = 1$ et 2).
3. On définit le *quantile* $t_{\alpha;k}$ de T par $P\left(T > t_{\alpha;k}\right) = \alpha$ 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 $N(0, 1)$ lorsque k est grand.

Calculs avec la loi de Student

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

1. Les quantiles $t_{\alpha;k}$ de la loi de Student sont données en annexe.
2. En utilisant R: $t_{\alpha;k}$ est donnée par `qt(1-α, k)`.
3. En utilisant Excel: `-LOI.STUDENT.INVERSE.N(α, k)`.

Exercice 5: Calculer $t_{0,9;3}$ et $P(X \leq 2,015)$ si $X \sim t_5$.

Utilisation de la loi de Student

Theorem 3.

Soit $X_1, X_2, ..., X_n$ un échantillon de taille n d'une variable aléatoire normale $X \sim N(\mu, \sigma^2)$. 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

$$T = \frac{\left(\overline{X} - \mu\right)}{\frac{S}{\sqrt{n}}}$$

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

Utilisation de la loi de Student (ii)

Exercice 6: 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 échantillonnale d'une différence de deux moyennes

Distribution d'une différence de moyennes

Considérons maintenant deux échantillons aléatoires indépendants $X_1, X_2, ..., X_{nX}$ et $Y_1, Y_2, ..., Y_{nY}$ 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 $\overline{X} - \overline{Y}$.

Theorem 4.

Dans la situation décrite ci-dessus:

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

Distribution d'une différence de moyennes (ii)

Theorem 5.

La variable aléatoire

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

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

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

Distribution d'une différence de moyennes (iii)

Exercice 6:
Soit $X \sim N(175, 10^2)$ et $Y \sim N(170, 9^2)$ 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 échantillonnale d'un rapport de variances

Distribution d'un rapport de variances

Considéreons 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 $N(\mu_X, \sigma_X^2)$ et $N(\mu_Y, \sigma_Y^2)$ respectivement.
On s'intéresse au rapport des variances échantillonnales $\frac{S_X^2}{S_Y^2}$.

Loi de Fisher

Definition 5.

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

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

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

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

Loi de Fisher (ii)

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

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

Propriété:

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

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

Calculs avec la loi de Fisher

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

1. Les quantiles de $F_{u,v}$ sont donnés en annexe.
2. En utilisant le logiciel R: $F_{\alpha;u,v}$ est donné par `qf(1-α, u, v)`.
3. En utilisant le logiciel Excel: $F_{\alpha;u,v}$ est donné par `INVERSE.LOI.F.N(1-α, u, v)`.

Exercice 7: Calculer $F_{0,75;11,10}$ et $P(X \leq 200)$ si $X \sim F_{2,1}$.

Distribution d'un rapport de variances (suite)

Theorem 6.

Soit $X_1, X_2, ..., X_{nX}$ et $Y_1, Y_2, ..., Y_{nY}$ deux échantillons aléatoire indépendants, de taille n_X et n_Y, des variables aléatoires X et Y.
On suppose que X et Y suivent des lois normales $N(\mu_X, \sigma_X^2)$ et $N(\mu_Y, \sigma_Y^2)$ respectivement.
Soit S_X^2 et S_Y^2 les variances échantillonnales. Alors la variable aléatoire

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

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

Distribution d'un rapport de variances (suite) (ii)

Exercice 7:
Soit $X \sim N(175, 10^2)$ et $Y \sim N(170, 9^2)$ la taille (en cm) des étudiants et étudiantes d'une école génie.
On choisit un échantillon de taille 50 de X et un échantillon de taille 25 de Y.
Quell est la probabilité que le rapport $\frac{S_X^2}{S_Y^2}$ soit inférieur à 3?