Statistiques inférentielles

Chapitre 3 : Échantillonnage

Julian Tugaut

Télécom Saint-Étienne

2015

Sommaire

Plan

Définitions

On se donne une population Ω de taille N avec N grand.

Définitions

On se donne une population Ω de taille N avec N grand.

Définition : Échantillon

On appelle échantillon un sous-ensemble de la population Ω.

Définitions

On se donne une population Ω de taille N avec N grand.

Définition : Échantillon

On appelle échantillon un sous-ensemble de la population Ω.

Définition : Taille de l'échantillon

Un échantillon de taille n est une liste de n individus $(\omega_1, \cdots, \omega_n)$ extraits de la population Ω.

Définitions

On se donne une population Ω de taille N avec N grand.

Définition : Échantillon

On appelle échantillon un sous-ensemble de la population Ω.

Définition : Taille de l'échantillon

Un échantillon de taille n est une liste de n individus $(\omega_1, \cdots, \omega_n)$ extraits de la population Ω.

On dit que Ω est la population mère de l'échantillon.

Définitions - 2

Exemple

On considère une population constituée de cinquante étudiants. On a ainsi $\Omega = \{\omega_1, \cdots, \omega_{50}\}$ et $N = 50$. On prend un échantillon de six étudiants, $\{\omega_4, \omega_8, \omega_{15}, \omega_{16}, \omega_{23}, \omega_{42}\}$ et $n = 6$.

Définitions - 2

Exemple

On considère une population constituée de cinquante étudiants. On a ainsi $\Omega = \{\omega_1, \cdots, \omega_{50}\}$ et $N = 50$. On prend un échantillon de six étudiants, $\{\omega_4, \omega_8, \omega_{15}, \omega_{16}, \omega_{23}, \omega_{42}\}$ et $n = 6$.

Définition : Échantillonnage

On appelle échantillonnage le prélèvement d'échantillons.

Définitions - 3

Définition : Taux d'échantillonnage

Le rapport t de l'effectif n de l'échantillon sur l'effectif N de la population mère dans laquelle il a été prélevé, est appelé taux d'échantillonnage : $t = \frac{n}{N}$.

Définitions - 3

Définition : Taux d'échantillonnage

Le rapport t de l'effectif n de l'échantillon sur l'effectif N de la population mère dans laquelle il a été prélevé, est appelé taux d'échantillonnage : $t = \frac{n}{N}$.

On parle aussi de fraction de sondage.

Définitions - 3

Définition : Taux d'échantillonnage

Le rapport t de l'effectif n de l'échantillon sur l'effectif N de la population mère dans laquelle il a été prélevé, est appelé taux d'échantillonnage : $t = \frac{n}{N}$.

On parle aussi de fraction de sondage.

Exemple

Dans l'Exemple, le taux d'échantillonnage est égal à $\frac{6}{50} = 0.12 = 12\%$.

Définitions - 3

Définition : Taux d'échantillonnage

Le rapport t de l'effectif n de l'échantillon sur l'effectif N de la population mère dans laquelle il a été prélevé, est appelé taux d'échantillonnage : $t = \frac{n}{N}$.

On parle aussi de fraction de sondage.

Exemple

Dans l'Exemple, le taux d'échantillonnage est égal à $\frac{6}{50} = 0.12 = 12\%$.

Définition

On appelle échantillonnage aléatoire un prélèvement de n individus dans une population mère tel que toutes les combinaisons possibles de n individus aient la même probabilité d'être prélevés.

But de l'échantillonnage

On cherche à décrire un caractère qualitatif ou quantitatif (discret ou continu) C dans une population mère Ω à travers l'étude des résultats obtenus sur un échantillon de taille n.

But de l'échantillonnage

On cherche à décrire un caractère qualitatif ou quantitatif (discret ou continu) C dans une population mère Ω à travers l'étude des résultats obtenus sur un échantillon de taille n.

Remarque

Le taux d'échantillonnage doit répondre à deux critères : il faut qu'il soit suffisamment élevé pour rendre compte de la population mère et il faut qu'il soit suffisamment petit pour être simple à étudier.

Exemple

Exemple

Étant donnée une population d'étudiants, on peut s'intéresser à un caractère quantitatif discret comme la note à un partiel (0, 5, 15, 20...).

Exemple

Exemple

Étant donnée une population d'étudiants, on peut s'intéresser à un caractère quantitatif discret comme la note à un partiel (0, 5, 15, 20...).
Étant donnée une population d'étudiants, on peut s'intéresser à un caractère quantitatif continu comme la taille.

Exemple

Exemple

Étant donnée une population d'étudiants, on peut s'intéresser à un caractère quantitatif discret comme la note à un partiel (0, 5, 15, 20...).
Étant donnée une population d'étudiants, on peut s'intéresser à un caractère quantitatif continu comme la taille.
Étant donnée une population d'étudiants, on peut s'intéresser à un caractère qualitatif comme la matière préférée ("probabilités", "signaux et systèmes discrets", "statistiques", "signaux aléatoires", "probabilités approfondies", "la sieste"...).

n-échantillon d'une variable aléatoire

Définition : n-échantillon de valeurs de X

Soit C un caractère quantitatif défini sur une population mère Ω. C est la réalisation d'une variable aléatoire X définie sur Ω par $X(\omega_i) = x_i$.

n-échantillon d'une variable aléatoire

Définition : n-échantillon de valeurs de X

Soit C un caractère quantitatif défini sur une population mère Ω.
C est la réalisation d'une variable aléatoire X définie sur Ω par $X(\omega_i) = x_i$.
On appelle n-échantillon de valeurs de X la liste des valeurs $(x_1, \cdots, x_n)$ observées prises par X sur un échantillon $(\omega_1, \cdots, \omega_n)$ de la population Ω.

Définition : n-échantillon de valeurs de X

Soit C un caractère quantitatif défini sur une population mère Ω.
C est la réalisation d'une variable aléatoire X définie sur Ω par $X(\omega_i) = x_i$.
On appelle n-échantillon de valeurs de X la liste des valeurs $(x_1, \cdots, x_n)$ observées prises par X sur un échantillon $(\omega_1, \cdots, \omega_n)$ de la population Ω.
Les coordonnées peuvent être considérées comme les valeurs des réalisations d'un vecteur de variables aléatoires $(X_1, \cdots, X_n)$ appelé n-échantillon de X où les variables aléatoires réelles X_i sont indépendantes et identiquement distribuées (de même loi).

n-échantillon d'une variable aléatoire

Définition : n-échantillon de valeurs de X

Soit C un caractère quantitatif défini sur une population mère Ω. C est la réalisation d'une variable aléatoire X définie sur Ω par $X(\omega_i) = x_i$.
On appelle n-échantillon de valeurs de X la liste des valeurs $(x_1, \cdots, x_n)$ observées prises par X sur un échantillon $(\omega_1, \cdots, \omega_n)$ de la population Ω.
Les coordonnées peuvent être considérées comme les valeurs des réalisations d'un vecteur de variables aléatoires $(X_1, \cdots, X_n)$ appelé n-échantillon de X où les variables aléatoires réelles X_i sont indépendantes et identiquement distribuées (de même loi).

X_1 est alors la variable aléatoire “valeur du premier élément de l'échantillon”, X_2 la variable aléatoire “valeur du deuxième élément de l'échantillon”, etc.

Plan

Définition

Définition : Statistique

On appelle statistique toute variable aléatoire qui s'écrit à l'aide des variables aléatoires $X_1, \cdots, X_n$.

Définition

Définition : Statistique

On appelle statistique toute variable aléatoire qui s'écrit à l'aide des variables aléatoires $X_1, \cdots, X_n$.

Remarque

Une statistique est donc une variable aléatoire qui est mesurable par rapport à la tribu engendrée par les n variables aléatoires $X_1, \cdots, X_n$.

Exemples

Exemple

La variable aléatoire

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

est une statistique.

Exemples

Exemple

La variable aléatoire

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

est une statistique.

Exemple

La variable aléatoire

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2$$

est aussi une statistique.

Exemples

Exemple

La variable aléatoire

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

est une statistique.

Exemple

La variable aléatoire

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2$$

est aussi une statistique.

Contre-exemple

La variable aléatoire X_{n+1} n'est pas une statistique.

Plan

Moyenne d'échantillon, variance d'échantillon

Définition : Moyenne d'échantillon

On définit la variable aléatoire $\overline{X_n}$, appelée moyenne d'échantillon par :

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i \, .$$

Moyenne d'échantillon, variance d'échantillon

Définition : Moyenne d'échantillon

On définit la variable aléatoire $\overline{X_n}$, appelée moyenne d'échantillon par :

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i \,.$$

Définition : Variance d'échantillon

On définit la variable aléatoire S_n^2, appelée variance d'échantillon par :

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X_n}\right)^2 \,.$$

Distribution d'échantillonnage de la moyenne - 1

Théorème

Quelle que soit la loi de la variable aléatoire X d'espérance μ et d'écart-type σ et quel que soit n, on a

$$\mathbb{E}\left[\overline{X_n}\right] = \mu = \mathbb{E}[X]\,,$$

Distribution d'échantillonnage de la moyenne - 1

Théorème

Quelle que soit la loi de la variable aléatoire X d'espérance μ et d'écart-type σ et quel que soit n, on a

$$\mathbb{E}\left[\overline{X_n}\right] = \mu = \mathbb{E}[X]\,,$$

ainsi que

$$\mathrm{Var}\left[\overline{X_n}\right] = \frac{\sigma^2}{n} = \frac{\mathrm{Var}[X]}{n}\,.$$

Distribution d'échantillonnage de la moyenne - 2

Preuve

La linéarité de l'espérance nous donne

$$\begin{aligned}\mathbb{E}\left[\overline{X_n}\right] &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X]\\ &= \mathbb{E}[X]\,,\end{aligned}$$

ce qui achève la première partie de la preuve.

Distribution d'échantillonnage de la moyenne - 3

On calcule désormais la variance de la moyenne d'échantillon. Par définition de la variance, il vient

$$\mathrm{Var}\left[\overline{X_n}\right] = \mathbb{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right)^2\right\}.$$

Or, nous venons de voir $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$. Ainsi :

$$\begin{aligned}
\mathrm{Var}\left[\overline{X_n}\right] &= \mathbb{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right\} \\
&= \mathbb{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)\right)^2\right\} \\
&= \frac{1}{n^2}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} (X_i - \mu)(X_j - \mu)\right\}.
\end{aligned}$$

Distribution d'échantillonnage de la moyenne - 4

La linéarité de l'espérance nous donne immédiatement

$$\mathrm{Var}\left[\overline{X_n}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\{(X_i-\mu)\left(X_j-\mu\right)\right\}\,.$$

On se sert maintenant de l'hypothèse d'indépendance des variables aléatoires X_i et X_j si $i \neq j$ pour obtenir :

$$\begin{aligned}\mathbb{E}\left\{(X_i-\mu)\left(X_j-\mu\right)\right\} &= \mathbb{E}\left(X_i-\mu\right)\times\mathbb{E}\left(X_j-\mu\right)\\ &= \left(\mathbb{E}[X_i]-\mu\right)\times\left(\mathbb{E}[X_j]-\mu\right)\,.\end{aligned}$$

Or, X_i et X_j ont même loi que X d'où $\mathbb{E}[X_i]-\mu = \mathbb{E}[X_j]-\mu = 0$.

Distribution d'échantillonnage de la moyenne - 5

Par conséquent, on obtient

$$\begin{aligned}
\mathrm{Var}\left[\overline{X_n}\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\{(X_i-\mu)^2\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X] \\
&= \frac{1}{n^2}\times n\times\mathrm{Var}[X] \\
&= \frac{\mathrm{Var}[X]}{n}\,.
\end{aligned}$$

Distribution d'échantillonnage de la moyenne - 6

Propriété

La variable aléatoire $\overline{X_n}$ converge en probabilité vers $\mu = \mathbb{E}[X]$.

Distribution d'échantillonnage de la moyenne - 6

Propriété

La variable aléatoire $\overline{X_n}$ converge en probabilité vers $\mu = \mathbb{E}[X]$.

Preuve

Par définition, la convergence en probabilité de $\overline{X_n}$ vers μ signifie que pour tout $\epsilon > 0$, on a

$$\lim_{n \to \infty} \mathbb{P}\left(\left|\overline{X_n} - \mu\right| > \epsilon\right) = 0\,.$$

On utilise l'inégalité de Bieanymé-Tchebychev :

$$\mathbb{P}\left(\left|\overline{X_n} - \mu\right| > \epsilon\right) \leq \frac{\mathrm{Var}\left[\overline{X_n}\right]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}\frac{1}{n} \longrightarrow 0\,.$$

La preuve est ainsi achevée.

Distribution d'échantillonnage de la variance

Théorème

Quelle que soit la loi de la variable aléatoire X d'espérance μ et d'écart-type σ et quel que soit n, on a

$$\mathbb{E}\left[S_n^2\right] = \frac{n-1}{n}\sigma^2 = \frac{n-1}{n}\mathrm{Var}[X]\,.$$

Distribution d'échantillonnage de la variance - 2

Preuve

$$
\begin{aligned}
S_n^2 :=& \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}\left\{(X_i-\mu) - \left(\overline{X_n}-\mu\right)\right\}^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\overline{X_n}-\mu\right)^2 - \frac{2}{n}\left(\overline{X_n}-\mu\right)\sum_{i=1}^{n}(X_i-\mu) \\
=& \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \left(\overline{X_n}-\mu\right)^2 .
\end{aligned}
$$

La linéarité de l'espérance donne alors

$$
\mathbb{E}\left[S_n^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{(X_i-\mu)^2\right\} - \mathbb{E}\left\{\left(\overline{X_n}-\mu\right)^2\right\} .
$$

Distribution d'échantillonnage de la variance - 3

Le second terme est égal à la variance de $\overline{X_n}$. Conséquemment, il vient

$$\begin{aligned}\mathbb{E}\left[S_n^2\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X_i] - \mathrm{Var}\left[\overline{X_n}\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X] - \frac{\sigma^2}{n}\\ &= \mathrm{Var}[X] - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2\,,\end{aligned}$$

ce qui achève la preuve du théorème.

Distribution d'échantillonnage de la variance - 4

Le calcul de la variance de la variance d'échantillonnage nécessite que la variable aléatoire X admette un moment d'ordre quatre fini.

Distribution d'échantillonnage de la variance - 4

Le calcul de la variance de la variance d'échantillonnage nécessite que la variable aléatoire X admette un moment d'ordre quatre fini.

Propriété

Quelle que soit la loi de la variable aléatoire X d'espérance μ, d'écart-type σ et de moment centré d'ordre quatre égal à μ_4 (c'est-à-dire tel que $\mathbb{E}\left[(X-\mu)^4\right] = \mu_4$) et quel que soit n, on a

$$\mathrm{Var}\left[S_n^2\right] = \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\sigma^4\right) .$$

Distribution d'échantillonnage de la variance - 5

Remarque

On dispose de l'équivalence $\mathrm{Var}\left[S_n^2\right] \sim \frac{\mu_4-\sigma^4}{n}$ pour n grand. Conséquemment, l'espérance de S_n^2 tend vers la variance de X et la variance de S_n^2 tend vers 0. On peut ainsi montrer la convergence en probabilité de S_n^2 vers $\sigma^2 = \mathrm{Var}[X]$.

Variance d'échantillon corrigée

Définition : Variance d'échantillon corrigée

On définit la variable aléatoire S_{n-1}^2, appelée variance d'échantillon corrigée, par

$$S_{n-1}^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2 .$$

Variance d'échantillon corrigée

Définition : Variance d'échantillon corrigée

On définit la variable aléatoire S_{n-1}^2, appelée variance d'échantillon corrigée, par

$$S_{n-1}^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2 .$$

On a donc

$$S_{n-1}^2 = \frac{n}{n-1} S_n^2 .$$

Conséquemment, il vient immédiatement :

Variance d'échantillon corrigée

Définition : Variance d'échantillon corrigée

On définit la variable aléatoire S_{n-1}^2, appelée variance d'échantillon corrigée, par

$$S_{n-1}^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X_n} \right)^2 .$$

On a donc

$$S_{n-1}^2 = \frac{n}{n-1} S_n^2 .$$

Conséquemment, il vient immédiatement :

Propriété

Quelle que soit la loi de la variable aléatoire X d'espérance μ et d'écart-type σ et quel que soit n, on a

$$\mathbb{E}\left[S_{n-1}^2 \right] = \sigma^2 = \mathrm{Var}[X] .$$

Plan

Moyenne d'échantillon

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}(\mu; \sigma)$, alors la moyenne empirique $\overline{X_n}$ suit la loi normale $\mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{N}}\right)$. Puis, la variable aléatoire $\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{N}}}$ suit la loi normale centrée réduite.

Moyenne d'échantillon

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}(\mu;\sigma)$, alors la moyenne empirique $\overline{X_n}$ suit la loi normale $\mathcal{N}\left(\mu;\frac{\sigma}{\sqrt{N}}\right)$. Puis, la variable aléatoire $\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{N}}}$ suit la loi normale centrée réduite.

Preuve

Comme les X_i ont même loi que X, les variables aléatoires X_i suivent la loi normale $\mathcal{N}(\mu;\sigma)$. Conséquemment, la variable aléatoire $\overline{X_n}$ suit une loi normale. Or, d'après le Théorème, son espérance vaut μ et sa variance vaut $\frac{\sigma^2}{n}$ ce qui achève de prouver que $\overline{X_n}$ suit la loi normale $\mathcal{N}\left(\mu;\frac{\sigma}{\sqrt{n}}\right)$.
Puis, en lui retranchant son espérance et en la divisant par son écart-type, on obtient une variable aléatoire qui suit la loi normale centrée réduite.

Variance d'échantillon - 1

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}(\mu;\sigma)$, alors la variable aléatoire $(n-1)\frac{S_{n-1}^2}{\sigma^2}$ suit la loi du Khi-deux à $n-1$ degrés de liberté, $\chi^2(n-1)$.

Variance d'échantillon - 1

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}(\mu;\sigma)$, alors la variable aléatoire $(n-1)\frac{S_{n-1}^2}{\sigma^2}$ suit la loi du Khi-deux à $n-1$ degrés de liberté, $\chi^2(n-1)$.

Le désavantage du Théorème est de ne pouvoir être appliqué que si l'on connaît la variance. Or, quand l'espérance est inconnue, la variance l'est également. On doit donc remplacer σ par son estimation, S_{n-1}^2 et considérer la statistique

$$\frac{\overline{X_n}-\mu}{\sqrt{\frac{S_{n-1}^2}{n}}}\,.$$

Variance d'échantillon - 2

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}(\mu;\sigma)$, alors la variable aléatoire $\frac{\overline{X_n}-\mu}{\sqrt{\frac{S_{n-1}^2}{n}}}$ suit la loi de Student à $n-1$ degrés de liberté, $\mathcal{T}(n-1)$.

Variance d'échantillon - 2

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}(\mu;\sigma)$, alors la variable aléatoire $\frac{\overline{X_n}-\mu}{\sqrt{\frac{S_{n-1}^2}{n}}}$ suit la loi de Student à $n-1$ degrés de liberté, $\mathcal{T}(n-1)$.

D'après les Théorèmes, il suffit ensuite de disposer des tables de la loi du Khi-deux et de la loi de Student.

Plan

Théorème central de la limite

Théorème

Soit $(X_i)_i$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées. On suppose $\mathbb{E}[X_1] = \mu$ et $\mathrm{Var}[X_1] = \sigma^2$. Alors on a la convergence en loi de la variable aléatoire

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

vers la loi normale centrée réduite.

Moyenne d'échantillon

Théorème

Si X suit une loi quelconque et si n est suffisamment grand (en pratique, $n \geq 30$), alors $\overline{X_n}$ suit approximativement la loi $\mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$ donc $\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement la loi normale centrée réduite $\mathcal{N}(0; 1)$.

De plus, $\frac{\overline{X_n}-\mu}{\sqrt{\frac{S^2_{n-1}}{n}}}$ suit approximativement la loi normale centrée réduite $\mathcal{N}(0; 1)$.

Moyenne d'échantillon

Théorème

Si X suit une loi quelconque et si n est suffisamment grand (en pratique, $n \geq 30$), alors $\overline{X_n}$ suit approximativement la loi $\mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$ donc $\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement la loi normale centrée réduite $\mathcal{N}(0;1)$.

De plus, $\frac{\overline{X_n}-\mu}{\sqrt{\frac{S_{n-1}^2}{n}}}$ suit approximativement la loi normale centrée réduite $\mathcal{N}(0;1)$.

Remarque

Si X suit une loi quelconque et si $n < 30$, alors la loi de $\overline{X_n}$ est inconnue.

Plan

Définition

Il arrive que le caractère C à estimer ne soit pas quantitatif mais qualitatif. Soit p la proportion d'individus présentant le caractère étudié dans la population mère Ω.

Définition

Il arrive que le caractère C à estimer ne soit pas quantitatif mais qualitatif. Soit p la proportion d'individus présentant le caractère étudié dans la population mère Ω.

Définition

On définit la variable aléatoire F_n, appelée proportion d'échantillon ou fréquence statistique par

$$F_n := \frac{K_n}{n} \,,$$

où K_n est la variable aléatoire qui compte le nombre d'apparitions du caractère considéré dans un échantillon de taille n.

Distribution d'échantillonnage des proportions

Théorème

Pour tout $n \in \mathbb{N}^*$, on a

$$\mathbb{E}[F_n] = p\,,$$

Distribution d'échantillonnage des proportions

Théorème

Pour tout $n \in \mathbb{N}^*$, on a

$$\mathbb{E}[F_n] = p \,,$$

ainsi que

$$\mathrm{Var}[F_n] = \frac{p(1-p)}{n} \,.$$

Ainsi, la variance de la variable aléatoire réelle F_n tend vers 0. Il s'ensuit que F_n converge en probabilité vers p.

Approximation par la loi normale

Théorème

Si n est suffisamment grand ($n \geq 30$ en pratique), alors F_n suit approximativement la loi normale d'espérance p et d'écart-type $\sqrt{\frac{p(1-p)}{n}}$, $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$.

Exemple

Exemple

Selon une étude sur le comportement des gens, 25% d'entre eux ont déjà téléchargé illégalement "Game of Thrones". Si on interroge cent personnes prises au hasard, quelle est la probabilité qu'au moins 35 d'entre eux aient déjà téléchargé illégalement "Game of Thrones"?

Exemple

Exemple

Selon une étude sur le comportement des gens, 25% d'entre eux ont déjà téléchargé illégalement "Game of Thrones". Si on interroge cent personnes prises au hasard, quelle est la probabilité qu'au moins 35 d'entre eux aient déjà téléchargé illégalement "Game of Thrones"? La proportion d'échantillon F_{100} suit approximativement la loi normale d'espérance 0.25 et d'écart-type $\frac{\sqrt{0.25\times 0.75}}{10}$,
$\mathcal{N}\left(0.25; \frac{\sqrt{0.25\times 0.75}}{10}\right)$.

Exemple

Exemple

Selon une étude sur le comportement des gens, 25% d'entre eux ont déjà téléchargé illégalement "Game of Thrones". Si on interroge cent personnes prises au hasard, quelle est la probabilité qu'au moins 35 d'entre eux aient déjà téléchargé illégalement "Game of Thrones"? La proportion d'échantillon F_{100} suit approximativement la loi normale d'espérance 0.25 et d'écart-type $\frac{\sqrt{0.25\times 0.75}}{10}$, $\mathcal{N}\left(0.25; \frac{\sqrt{0.25\times 0.75}}{10}\right)$. On a alors

$$p = \mathbb{P}\left(F_{100} \geq 35\right) = \mathbb{P}\left(\frac{F_{100} - 0.25}{\frac{\sqrt{0.25\times 0.75}}{10}} \geq \frac{0.1}{\frac{\sqrt{0.25\times 0.75}}{10}}\right)$$

$$= \mathbb{P}\left(\frac{F_{100} - 0.25}{\frac{\sqrt{0.25\times 0.75}}{10}} \geq \frac{1}{\sqrt{0.25 \times 0.75}}\right).$$

Exemple

Exemple

Selon une étude sur le comportement des gens, 25% d'entre eux ont déjà téléchargé illégalement "Game of Thrones". Si on interroge cent personnes prises au hasard, quelle est la probabilité qu'au moins 35 d'entre eux aient déjà téléchargé illégalement "Game of Thrones"? La proportion d'échantillon F_{100} suit approximativement la loi normale d'espérance 0.25 et d'écart-type $\frac{\sqrt{0.25\times 0.75}}{10}$, $\mathcal{N}\left(0.25; \frac{\sqrt{0.25\times 0.75}}{10}\right)$. On a alors

$$
\begin{aligned}
p = \mathbb{P}\left(F_{100} \geq 35\right) &= \mathbb{P}\left(\frac{F_{100} - 0.25}{\frac{\sqrt{0.25\times 0.75}}{10}} \geq \frac{0.1}{\frac{\sqrt{0.25\times 0.75}}{10}}\right) \\
&= \mathbb{P}\left(\frac{F_{100} - 0.25}{\frac{\sqrt{0.25\times 0.75}}{10}} \geq \frac{1}{\sqrt{0.25 \times 0.75}}\right).
\end{aligned}
$$

On obtient alors $p \approx 1 - \Phi\left(\frac{1}{\sqrt{0.25\times 0.75}}\right) \approx 1 - \Phi(2.31) \approx 0.01044.$