10. Estimation

MTH2302D

S. Le Digabel, École Polytechnique de Montréal

A2017

(v2)

Plan

1. Introduction

2. Estimation ponctuelle

3. Estimation par intervalles de confiance

4. Autres problèmes d'estimation par intervalle de confiance

Introduction

L'*inférence statistique* consiste à tirer des conclusions sur une population à partir d'un échantillon.

Deux parties :

- Estimation de paramètres.
- Tests d'hypothèses.

Estimation de paramètres : deux méthodes :

- Estimation ponctuelle.
- Estimation par intervalles de confiance.

Estimation ponctuelle

But : estimer un paramètre d'une population à l'aide d'une statistique.

Définition

Soit X une variable aléatoire dont la distribution dépend d'un paramètre θ. Soit $X_1, \ldots, X_n$ un échantillon aléatoire de X de taille n.

Un *estimateur ponctuel* de θ est une statistique $\hat{\theta}$ de la forme $\hat{\theta} = h(X_1, \ldots, X_n)$ et vérifiant certains critères.

Exemple 1

Soit X une variable aléatoire et $X_1, \ldots, X_n$ un échantillon de taille n. Alors

- $\hat{\theta}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $\hat{\theta}_2 = X_i$ (une valeur individuelle)

sont des estimateurs de la moyenne $\theta = \mu = \mathsf{E}(X)$.

Exemple 2

Soit X une variable aléatoire et $X_1, \ldots, X_n$ un échantillon de taille n. Alors

- $\hat{\theta}_1 = S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2$
- $\hat{\theta}_2 = S_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2$

sont des estimateurs de la variance $\theta = \sigma^2 = \mathsf{V}(X)$.

Trois critères pour la qualité d'un estimateur

Critère 1 : le biais

Le *biais* d'un estimateur $\hat{\theta}$ du paramètre θ est

$$\mathsf{Biais}(\hat{\theta}) = \mathsf{E}(\hat{\theta}) - \theta.$$

On dit que $\hat{\theta}$ est *sans biais* ou *non-biaisé* si $\mathsf{Biais}(\hat{\theta}) = 0$.

Le biais est une mesure de l'erreur systématique faite en approximant θ par $\hat{\theta}$.

Exemple 3

Prouver que $\mathsf{E}(\overline{X}) = \mu$ et $\mathsf{E}(S^2) = \sigma^2$ et donc que $\overline{X}$ et S^2 sont des estimateurs sans biais de μ et σ^2.

Trois critères pour la qualité d'un estimateur (suite)

Critère 2 : Erreur quadratique moyenne

Définition

L'*erreur quadratique moyenne* (EQM) d'un estimateur $\hat{\theta}$ du paramètre θ est

$$\mathsf{EQM}(\hat{\theta}) = \mathsf{E}\left[(\hat{\theta} - \theta)^2\right].$$

L'EQM est une mesure de la précision d'un estimateur.

Théorème

Si $\hat{\theta}$ est un estimateur du paramètre θ alors

$$\mathsf{EQM}(\hat{\theta}) = \mathsf{V}(\hat{\theta}) + \left[\mathsf{Biais}(\hat{\theta})\right]^2.$$

Trois critères pour la qualité d'un estimateur (suite)

Critère 2 : Erreur quadratique moyenne (suite)

Le meilleur de deux estimateur $\hat{\theta}_1$ et $\hat{\theta}_2$, c'est-à-dire le plus *efficace*, est celui qui a la plus petite EQM : $\hat{\theta}_1$ est plus efficace que $\hat{\theta}_2$ si

$$\mathsf{EQM}(\hat{\theta}_1) < \mathsf{EQM}(\hat{\theta}_2) \quad \Leftrightarrow \quad \frac{\mathsf{EQM}(\hat{\theta}_1)}{\mathsf{EQM}(\hat{\theta}_2)} < 1.$$

Lorsque deux estimateurs son non biaisés, ceci revient à dire que le plus efficace est celui dont la variance est la plus petite.

Exemple 4

Soit $X_1, X_2, \ldots, X_5$ un échantillon aléatoire d'une v.a. X telle que $\mathsf{E}(X) = \mu$ et $\mathsf{V}(X) = \sigma^2$. Pour estimer μ, on considère

$$\hat{\theta_1} = \frac{X_1 + \ldots + X_5}{5} \text{ et } \hat{\theta_2} = \frac{2X_1 - X_2 + X_4}{2} \, .$$

1. Ces deux estimateurs sont-ils non-biaisés ?
2. Quel est le meilleur des deux ?

Trois critères pour la qualité d'un estimateur (suite)

Critère 3 : Convergence

Dénotons par $\hat{\theta}_n$ un estimateur du paramètre θ calculé à partir d'un échantillon de taille n.

Définition

Un estimateur $\hat{\theta}_n$ est *convergent* si pour tout $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta| < \varepsilon\right) = 1.$$

Ceci signifie : si la taille de l'échantillon est assez grande alors on est (presque) certain que l'estimateur $\hat{\theta}_n$ est très proche de θ.

Théorème

Si EQM$(\hat{\theta}_n)$ converge vers 0 lorsque $n \to \infty$ alors $\hat{\theta}_n$ est convergent.

Méthode des moments

- La plupart des lois que nous avons vues sont déterminées par un ou deux paramètres généralement liés aux deux premiers moments de la v.a., $\mu'_1 = \mu$ et $\mu'_2 = \sigma^2$.
- Soit $X \sim Loi(\theta_1, \theta_2)$ avec θ_1 et θ_2 inconnus mais dépendants des deux premiers moments. Si $X_1, X_2, \ldots, X_n$ est un échantillon de taille n des valeurs de X, on peut définir les deux premiers moments de l'échantillon par rapport à l'origine :

  $$m'_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \text{ avec } k \in \{1, 2\} \ .$$

  Ainsi on peut estimer μ par $\hat{\mu} = m'_1$ et σ^2 par $\hat{\sigma}^2 = m'_2 - (m'_1)^2$, et donc θ_1 et θ_2.

Exemple 5

Soit $X \sim$ Unif$(0, a)$. Quel est l'estimateur $\hat{a}$ du paramètre a par la méthode des moments ?

Méthode du maximum de vraisemblance

Soit X une variable aléatoire dont la distribution est donnée par $f(x, \theta)$, où θ est un paramètre inconnu.

Soit $x_1, \ldots, x_n$ une réalisation (valeurs observées) d'un échantillon aléatoire de taille n de X.

Définition

La *fonction de vraisemblance* de cet échantillon est

$$L(\theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta).$$

Intuitivement, $L(\theta)$ est la probabilité d'observer les $x_1, x_2, \ldots, x_n$: $P(X_1 = x_1 \cap X_2 = x_2 \cap \ldots \cap X_n = x_n)$.

Définition

L'*estimateur de vraisemblance maximale* de θ est la valeur $\hat{\theta}$ pour laquelle $L(\theta)$ atteint son maximum.

Exemple 6

Soit $X \sim \text{Bern}(p)$. Quel est l'estimateur $\hat{p}$ du paramètre p par la méthode du maximum de vraisemblance ?

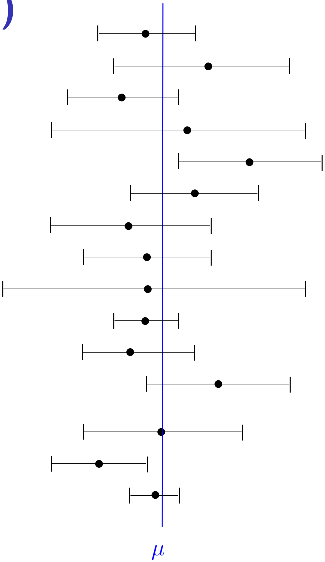
Voir exemple 10.5 page 235 (2ème édition) / exemple 9.5 page 248 (3ème édition) pour la loi normale.

Intervalles de confiance (IC)

Idée : Soit θ un paramètre de la distribution d'une variable aléatoire X.

À partir d'un échantillon, on cherche à déterminer un intervalle $[L, U]$ qui contient θ avec une probabilité donnée.

Sur ce dessin, par ex. :
$P(L \leq \mu \leq U) \simeq 87\%$.



Intervalles de confiance (suite)

Définition

Soit X une variable aléatoire et θ un paramètre de sa distribution.

Soit $X_1, \ldots, X_n$ un échantillon de taille n de X.

Si $L \equiv L(X_1, \ldots, X_n)$ et $U \equiv U(X_1, \ldots, X_n)$ sont deux statistiques telles que

$$P(L \leq \theta \leq U) = 1 - \alpha$$

alors on dit que $[L, U]$ est un *intervalle de confiance* pour θ de *niveau de confiance* $1 - \alpha$.

IC pour la moyenne μ : cas où σ^2 est connue

Rappel : si $X \sim \mathsf{N}(\mu, \sigma^2)$ ou si la taille n de l'échantillon est grand alors

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0, 1).$$

Théorème

Dans ce cas, l'intervalle de confiance à $100(1-\alpha)\%$ pour μ est

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

où $z_{\alpha/2}$ est un nombre tel que $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

Exemple 7

Prouver le théorème.

IC pour la moyenne μ : cas où σ^2 est connue (suite)

Remarques :

1. La valeur de $z_{\alpha/2}$ dépend du niveau de confiance voulu : par exemple, avec la table du site du cours, on a :
   - Si $1-\alpha = 0.90$, alors $z_{\alpha/2} \simeq 1.645$.
   - Si $1-\alpha = 0.95$, alors $z_{\alpha/2} \simeq 1.960$.
   - Si $1-\alpha = 0.99$, alors $z_{\alpha/2} \simeq 2.576$.
2. On peut aussi considérer un intervalle unilatéral, de la forme

$$[L, \infty[\quad \text{ou} \quad]-\infty, U]$$

correspondant à

$$P(\mu \geq L) = 1-\alpha \quad \text{ou} \quad P(\mu \leq U) = 1-\alpha.$$

Dans ce cas, on remplace $\alpha/2$ par α dans les bornes déjà trouvées :

$$L = \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \quad \text{et} \quad U = \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Exemple 8

Des tests sur la conductivité thermique d'un métal ont permis d'obtenir les données suivantes pour un échantillon de taille $n = 10$:

41.60	41.48	42.34	41.95	41.86
42.18	41.72	42.26	41.81	42.04

Soit X la conductivité thermique du métal. Sachant que X suit une loi normale d'écart-type $\sigma = 0.10$, donner :

1. Une estimation ponctuelle de $\mu = \mathsf{E}(X)$.
2. Un intervalle de confiance à 95% pour μ.
3. Un intervalle de confiance unilatéral, avec borne inférieure, au niveau de confiance 95% pour μ.

Niveau de confiance et précision de l'estimation

Soit X une variable aléatoire normale de variance connue pour laquelle on veut estimer la moyenne à l'aide d'un échantillon de taille n.

Si le niveau de confiance $1-\alpha$ augmente alors la longueur de l'intervalle de confiance augmente.

La plus grande différence $|\overline{X}-\mu|$ possible entre l'estimateur et le paramètre, appelée *erreur*, est égale à la moitié de la longueur de l'intervalle de confiance.

Par conséquent, si le niveau de confiance augmente, l'erreur augmente.

Taille de l'échantillon

Pour un niveau de confiance $1-\alpha$ donné, soit $|\overline{X}-\mu|$ l'erreur de l'estimation de μ par $\overline{X}$.

Si on exige que l'erreur soit inférieure à une valeur fixée E, quelle doit être la taille minimale de l'échantillon utilisé ?

Réponse :

$$n = \left\lceil \left(\frac{\sigma z_{\alpha/2}}{E}\right)^2 \right\rceil .$$

Exemple 9 : En faire la preuve.

Exemple 10

Des tests sur la conductivité thermique d'un métal ont permis d'obtenir les données suivantes pour un échantillon de taille $n = 10$.

41.60	41.48	42.34	41.95	41.86
42.18	41.72	42.26	41.81	42.04

Soit X la conductivité thermique du métal. Sachant que X suit une loi normale d'écart-type $\sigma = 0.10$, quelle taille d'échantillon est nécessaire pour construire un intervalle de confiance à 95% avec une erreur inférieure à 0.05 ?

Intervalle de confiance pour la moyenne : autres cas

Une procédure semblable au cas précédent (variance connue) permet de construire des intervalles de confiance pour la moyenne μ dans différentes situations, en utilisant les distributions échantillonnales étudiées auparavant :

- Cas où $X \sim \mathsf{N}(\mu, \sigma^2)$ et σ^2 est inconnue.
- Cas où n est très grand et σ^2 est inconnue.

IC pour μ : résumé

Situation	Relation utilisée	IC niveau $1-\alpha$
σ^2 est connue et $X \sim \mathsf{N}(\mu, \sigma^2)$, ou n grand	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1)$	$\mu = \overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
σ^2 est inconnue et $X \sim \mathsf{N}(\mu, \sigma^2)$	$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim \mathsf{T}_{n-1}$	$\mu = \overline{X} \pm t_{\alpha/2;n-1} \frac{S}{\sqrt{n}}$
σ^2 est inconnue et n est grand	$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim \mathsf{N}(0,1)$	$\mu = \overline{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$

Intervalles de confiance pour la variance

Une procédure semblable à celle pour la moyenne permet de construire des intervalles de confiance pour σ^2 dans différentes situations, en utilisant les distributions échantillonnales étudiées auparavant :

- Cas où μ est connue.
- Cas où $X \sim \mathsf{N}(\mu, \sigma^2)$ et μ inconnue.
- Cas où n est très grand et μ inconnue.

IC pour σ^2 et σ : résumé

Situation	Relation utilisée	IC niveau $1-\alpha$
μ est connue et $X \sim \mathrm{N}(\mu, \sigma^2)$	$n\frac{S_\mu^2}{\sigma^2} \sim \chi_n^2$ avec $S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$	$\frac{nS_\mu^2}{\chi^2_{\alpha/2;n}} \leq \sigma^2 \leq \frac{nS_\mu^2}{\chi^2_{1-\alpha/2;n}}$
μ inconnue et $X \sim \mathrm{N}(\mu, \sigma^2)$	$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$	$\frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}}$
μ inconnue et n grand	$\frac{S-\sigma}{\sigma/\sqrt{2n}} \sim \mathrm{N}(0,1)$	$\frac{S}{1+\frac{z_{\alpha/2}}{\sqrt{2n}}} \leq \sigma \leq \frac{S}{1-\frac{z_{\alpha/2}}{\sqrt{2n}}}$

Exemple 11

Afin d'estimer la variance σ^2 de l'épaisseur d'un certain type de verre, un échantillon de 25 spécimens est prélevé. L'écart-type observé dans l'échantillon est de 0.08 mm.

On suppose que l'épaisseur du verre est distribuée selon une loi normale.

1. Déterminer un intervalle de confiance à 95% pour σ^2.
2. Donner un intervalle de confiance unilatéral, avec borne supérieure, au niveau de confiance 90% pour σ^2.

Intervalle de confiance pour une proportion

Considérons une expérience aléatoire et soit p la proportion de succès dans une population. Soit X le nombre de succès dans un échantillon de très grande taille n. On a donc $X \sim \mathrm{B}(n, p)$ et $\hat{p} = \dfrac{X}{n} = \displaystyle\sum_{i=1}^{n} X_i/n$ est un estimateur pour p.

De plus, on a vu auparavant (approximation d'une binomiale par une normale), que si n est grand alors :

$$X \sim N(\mu = np, \sigma^2 = np(1-p)),$$

et donc $\frac{X-np}{\sqrt{np(1-p)}} \sim \mathrm{N}(0,1)$ et $\dfrac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathrm{N}(0,1)$.

IC pour une proportion (suite)

On a

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathrm{N}(0,1) \ .$$

Donc $P\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p}+z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha.$

Problème : p est inconnu et apparaît dans les bornes. Une approximation acceptable est de le remplacer par son estimateur $\hat{p}$, et ainsi l'IC est

$$p = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ .$$

IC pour une proportion : calcul de la taille de l'échantillon n

On veut borner l'erreur d'approximation
$|p - \hat{p}| = \left| z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right| \leq E$, ce qui donne

$$n \geq \left(\frac{z_{\alpha/2}}{E} \right)^2 \hat{p}(1 - \hat{p}) \ .$$

Comme en général on fait ce calcul avant de considérer un échantillon, on n'a pas forcément de valeur pour $\hat{p}$. Si on a une estimation antérieure $\hat{p_0}$, on la considère, sinon on prend $\hat{p} = 0.5$ et on calcule

$$n = \left\lceil \left(\frac{z_{\alpha/2}}{2E} \right)^2 \right\rceil \ .$$

Exemple 12

Douze des 75 arbres d'un échantillon aléatoire sont contaminés par une maladie. Pour le déterminer, on a du abattre ces arbres.

1. Déterminer un intervalle de confiance à 95% pour p la proportion d'arbres malades dans la forêt.
2. Après cette mesure, on veut une plus petite erreur d'approximation de p. Combien d'arbres supplémentaires abattre si on veut une erreur d'au plus 5% ?
3. Combien d'arbres on aurait du couper si on veut la même erreur maximale, mais si on n'avait pas considéré le premier échantillon de 75 individus ?

Intervalle de confiance avec deux échantillons

Une procédure semblable à celle pour la moyenne permet de construire des intervalles de confiance dans différentes situations où deux échantillons X_1 et X_2 sont obtenus, en utilisant les distributions échantillonnales étudiées auparavant :

- IC pour $\mu_1 - \mu_2$ lorsque $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ et les σ_i^2 sont inconnues mais égales.
- IC pour $\mu_1 - \mu_2$ lorsque $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ et les σ_i^2 sont inconnues et différentes.
- IC pour $\mu_1 - \mu_2$ lorsque n_1, n_2 sont grands et les σ_i^2 sont inconnues.
- Intervalle de confiance pour σ_1^2/σ_2^2 lorsque $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$.
- IC pour $p_1 - p_2$ lorsque n_1, n_2 sont grands.

IC pour $\mu_1 - \mu_2$

Situation	Relation utilisée	IC niveau $1-\alpha$
σ_1^2 et σ_2^2 connues $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ ou n_1, n_2 grands	$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathsf{N}(0,1)$	$\mu_1 - \mu_2 = \overline{X}_1 - \overline{X}_2$ $\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
σ_1^2, σ_2^2 inconnues $\sigma_1^2 = \sigma_2^2$ $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$	$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathsf{T}_{n_1+n_2-2}$ $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$	$\mu_1 - \mu_2 = \overline{X}_1 - \overline{X}_2$ $\pm t_{\alpha/2; n_1+n_2-2}$ $\times S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

IC pour $\mu_1 - \mu_2$ (suite)

Situation	Relation utilisée	IC niveau $1-\alpha$
σ_1^2, σ_2^2 inconnues $\sigma_1^2 \neq \sigma_2^2$ $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$	$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T_\nu$	$\mu_1 - \mu_2 = \overline{X}_1 - \overline{X}_2 \pm t_{\alpha/2;\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ $\nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2$
σ_1^2, σ_2^2 inconnues n_1, n_2 grands	$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathrm{N}(0,1)$	$\mu_1 - \mu_2 = \overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

IC pour $\mu_1 - \mu_2$: observations couplées

Supposons que les données aient été recueillies par paires sur les mêmes unités expérimentales, c'est-à-dire que chaque unité fournit deux observations X_1 et X_2.

Si X_1 et X_2 suivent des lois normales alors $D = X_1 - X_2$ suit une loi normale et

$$\frac{\overline{D} - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \sim \mathsf{T}_{n-1}$$

où $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \overline{X_1} - \overline{X_2}$ et $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$.

L'IC pour $\mu_1 - \mu_2$ est :

$$\boxed{\overline{D} - t_{\alpha/2;n-1}\frac{S_D}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \overline{D} + t_{\alpha/2;n-1}\frac{S_D}{\sqrt{n}}}$$

Exemple 13

On a calculé les valeurs suivantes pour les notes d'un groupe (échantillon) de 41 étudiants de MTH2302D pour le contrôle périodique et l'examen final :

Moyenne du contrôle sur 30	Moyenne du final sur 50	Écart-type des différences (sur 50)
17.45	31.75	6.48

Comme les deux examens ont été passés par les mêmes étudiants, on considère que les notes obtenues à ces deux évaluations sont des observations couplées.

Dans ce cas, déterminez un intervalle de confiance à 95% pour la différence des moyennes.

IC pour σ_1^2/σ_2^2 et $p_1 - p_2$

Situation	Relation utilisée	IC niveau $1-\alpha$
$X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$	$\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim F_{n_2-1,n_1-1}$	$L \leq \sigma_1^2/\sigma_2^2 \leq U$ avec $L = \frac{S_1^2}{S_2^2} F_{1-\alpha/2;n_2-1,n_1-1}$ $U = \frac{S_1^2}{S_2^2} F_{\alpha/2;n_2-1,n_1-1}$
X_i binomiale n_1, n_2 grands	$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$ $\sim \mathsf{N}(0,1)$	$p_1 - p_2 = \hat{p}_1 - \hat{p}_2$ $\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

1. Introduction

2. Estimation ponctuelle

3. Estimation par intervalles de confiance

4. Autres problèmes d'estimation par intervalle de confiance

Intervalle de prévision

Contexte : On tire un échantillon $X_1, \ldots, X_n$ d'une population normale $X \sim \mathrm{N}(\mu, \sigma^2)$. On veut prédire la prochaine observation X_{n+1}.

Définition

On construit un *intervalle de prévision* comme suit :

- Un estimateur ponctuel de X_{n+1} est $\overline{X}$.
- Puisque les v.a. sont indépendantes, $X_{n+1} - \overline{X}$ suit une loi normale de moyenne nulle et de variance $\sigma^2 + \frac{\sigma^2}{n}$.
- On construit l'intervalle de confiance correspondant.

Intervalle de prévision (suite)

- L'intervalle de prévision est

$$\overline{X} - z_{\alpha/2}\sqrt{\sigma^2\left(1+\frac{1}{n}\right)} \leq X_{n+1} \leq \overline{X} + z_{\alpha/2}\sqrt{\sigma^2\left(1+\frac{1}{n}\right)}$$

si la variance σ^2 est connue.

- L'intervalle de prévision est

$$\overline{X} - t_{\alpha/2;n-1}\sqrt{S^2\left(1+\frac{1}{n}\right)} \leq X_{n+1} \leq \overline{X} + t_{\alpha/2;n-1}\sqrt{S^2\left(1+\frac{1}{n}\right)}$$

si la variance σ^2 n'est pas connue.

Exemple 14

On a mesuré au cours de 10 vols l'accélération maximale (en g) d'un avion de ligne. Les résultats obtenus sont :

1.15 1.23 1.56 1.69 1.71 1.83 1.83 1.85 1.90 1.91

On veut prédire l'accélération maximale de l'avion lors de son prochain vol, en supposant que l'accélération maximale suit une loi normale, avec une confiance de 95%.

Intervalle de tolérance

Contexte : On veut construire, à partir d'un échantillon, un intervalle qui contient un pourcentage donné des valeurs de la population, avec une probabilité $1-\alpha$.

Définition

Un *intervalle de tolérance* d'une population X est un intervalle construit à partir d'un échantillon $X_1, \ldots, X_n$ et qui contient $q\%$ des valeurs de X avec probabilité $1-\alpha$.

- q est le *taux de couverture*.
- $1-\alpha$ est le *coefficient de confiance*.

Intervalle de tolérance avec une loi normale

On considère $X \sim \mathsf{N}(\mu, \sigma^2)$.

Si μ et σ^2 sont connues alors déterminer un intervalle de tolérance se réduit au calcul d'une probabilité avec la loi normale.

Si μ et σ^2 ne sont pas connues alors l'intervalle de tolérance est de la forme

$$\left[\, \overline{X} - kS; \overline{X} + kS \,\right]$$

où k est une constante dépendant de q et $1-\alpha$.

Les valeurs de k pour différentes combinaisons du taux de couverture q et du niveau de confiance $1-\alpha$ sont habituellement données dans des tables (voir page 537, 3ème édition, cas bilatéral).

Exemple 15

On a mesuré au cours de 10 vols l'accélération maximale (en g) d'un avion de ligne. Les résultats obtenus sont :

$$1.15 \quad 1.23 \quad 1.56 \quad 1.69 \quad 1.71 \quad 1.83 \quad 1.83 \quad 1.85 \quad 1.90 \quad 1.91$$

On veut définir un intervalle de tolérance bilatéral dont on peut être confiant à 95% qu'il comprend 99% de toutes les accélérations maximales possibles.

Intervalle de tolérance (cas général)

Ici X n'est pas normale, et l'intervalle de tolérance est $[X_{\min}; X_{\max}]$. La relation entre le taux de couverture et le niveau de confiance est

$$\alpha = nq^{n-1} - (n-1)q^n$$

où n est la taille de l'échantillon.

Deux situations sont possibles :

1. Si q et n sont connus alors on peut déterminer le niveau de confiance $1 - \alpha$.
2. Si α et q sont connus alors on peut déterminer n (approximativement) par

$$n \simeq \left\lceil \frac{1}{2} + \frac{1+q}{1-q} \, \frac{\chi^2_{\alpha;4}}{4} \right\rceil.$$

Exemple 16

On s'intéresse à la durée nécessaire pour accomplir une certaine tâche. On dispose d'un échantillon de 50 mesures de cette durée (en minutes). Une analyse statistique descriptive des 50 mesures montre que $X_{\min} = 8.3$ et $X_{\max} = 11.8$. De plus, les données ne semblent pas provenir d'une distribution normale.

1. Déterminer un intervalle de tolérance pour la durée X ainsi que le niveau de confiance si le taux de couverture est de 95%.
2. Quelle taille d'échantillon doit-on utiliser pour que le niveau de confiance soit de 95% et le taux de couverture de 95% ?