Statistique (MATH-F-315, Cours #4)

Thomas Verdebout

Université Libre de Bruxelles

2015

Plan de la partie Statistique du cours

1. Introduction.
2. Théorie de l'estimation.
3. Tests d'hypothèses et intervalles de confiance.
4. Régression.
5. ANOVA.

Intervalles de confiance

L'estimation ponctuelle ne fournit "qu'un estimateur" mais ne donne pas d'indication sur la précision de l'estimateur. Les statistiques de test obtenues pour résoudre des problèmes de test peuvent généralement être "inversées" pour obtenir ce que nous appellerons des intervalles de confiance.

Definition

Un *intervalle de confiance* au *niveau de confiance* $(1-\alpha)$ pour θ est un intervalle $[I^-(\mathbf{X}), I^+(\mathbf{X})]$ tel que

(i) $I^-(\mathbf{X})$ et $I^+(\mathbf{X})$ sont des statistiques;

(ii) $\mathrm{P}_\theta[I^-(\mathbf{X}) \leq \theta \leq I^+(\mathbf{X})] \geq 1-\alpha$ pour tout $\theta \in \Theta$.

IC pour la moyenne d'un échantillon gaussien

1) Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ ($n \geq 2$). Supposons que σ^2 est connu. On sait grâce au lemme de Fisher que:

$$\bar{X}^{(n)} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X}^{(n)} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}\left(0, 1\right)$$

$$\Downarrow$$

$$\begin{aligned}
\mathrm{P}_\mu\left[\frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] &= \alpha/2 \quad \text{pour tout } \mu \\
\mathrm{P}_\mu\left[\frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] &= 1-\alpha/2 \quad \text{pour tout } \mu \\
\mathrm{P}_\mu\left[\frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} \geq z_{1-\alpha/2}\right] &= \alpha/2 \quad \text{pour tout } \mu \\
\Longrightarrow \mathrm{P}_\mu\left[z_{\alpha/2} \leq \frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] &= 1-\alpha \quad \text{pour tout } \mu
\end{aligned}$$

Notons également que $z_{\alpha/2} = -z_{1-\alpha/2}$.

IC pour la moyenne d'un échantillon gaussien

$$
\begin{aligned}
\mathrm{P}_\mu\left[z_{\alpha/2} \leq \frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] &= 1-\alpha \quad \forall \mu \\
\mathrm{P}_\mu\left[z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X}^{(n)} - \mu \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] &= 1-\alpha \quad \forall \mu \\
\mathrm{P}_\mu\left[\bar{X}^{(n)} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}^{(n)} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] &= 1-\alpha \quad \forall \mu \\
\mathrm{P}_\mu\left[\underbrace{\bar{X}^{(n)} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}}_{=:l^-(\mathbf{X})} \leq \mu \leq \underbrace{\bar{X}^{(n)} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}}_{=:l^+(\mathbf{X})}\right] &= 1-\alpha \quad \forall \mu
\end{aligned}
$$

et $[l^-(\mathbf{X}), l^+(\mathbf{X})] = \left[\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance pour μ au niveau de confiance $(1-\alpha)$.

IC pour la moyenne d'un échantillon gaussien

2) Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ ($n \geq 2$) où $\sigma^2 = \mathrm{Var}[X_i] < \infty$ est un paramètre de nuisance. Grâce au lemme de Fisher:

$$\bar{X}^{(n)} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad \textit{et} \qquad \frac{ns^2}{\sigma^2} \sim \chi^2_{n-1},$$

il découle donc du lemme de Fisher que

$$\frac{\frac{\sqrt{n}\,(\bar{X}^{(n)} - \mu)}{\sigma}}{\sqrt{\left(\frac{ns^2}{\sigma^2}\right)/(n-1)}} = \frac{(\bar{X}^{(n)} - \mu)}{s/\sqrt{n-1}} = \frac{(\bar{X}^{(n)} - \mu)}{S/\sqrt{n}} \sim t_{n-1}.$$

$$\Downarrow$$

$$\mathrm{P}_{\mu,\sigma^2}\left[\frac{\bar{X}^{(n)} - \mu}{S/\sqrt{n}} \leq t_{n-1;\alpha/2}\right] = \alpha/2 \quad \text{pour tout } \mu, \sigma^2$$

$$\mathrm{P}_{\mu,\sigma^2}\left[\frac{\bar{X}^{(n)} - \mu}{S/\sqrt{n}} \geq t_{n-1;1-\alpha/2}\right] = \alpha/2 \quad \text{pour tout } \mu, \sigma^2$$

$$\Longrightarrow \mathrm{P}_{\mu,\sigma^2}\left[t_{n-1;\alpha/2} \leq \frac{\bar{X}^{(n)} - \mu}{S/\sqrt{n}} \leq t_{n-1;1-\alpha/2}\right] = 1 - \alpha \quad \text{pour tout } \mu, \sigma^2$$

Notons également que $t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$.

IC pour la moyenne d'un échantillon gaussien

$$\begin{aligned}
\mathrm{P}_{\mu,\sigma^2}\left[t_{n-1;\alpha/2} \leq \frac{\bar{X}^{(n)}-\mu}{S/\sqrt{n}} \leq t_{n-1;1-\alpha/2}\right] &= 1-\alpha \quad \forall \mu, \sigma^2 \\
\mathrm{P}_{\mu,\sigma^2}\left[t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \leq \bar{X}^{(n)}-\mu \leq t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right] &= 1-\alpha \quad \forall \mu, \sigma^2 \\
\mathrm{P}_{\mu,\sigma^2}\left[\bar{X}^{(n)}-t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}^{(n)}-t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right] &= 1-\alpha \quad \forall \mu, \sigma^2 \\
\mathrm{P}_{\mu,\sigma^2}\left[\underbrace{\bar{X}^{(n)}-t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}}_{=:I^-(\mathbf{X})} \leq \mu \leq \underbrace{\bar{X}^{(n)}+t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}}_{=:I^+(\mathbf{X})}\right] &= 1-\alpha \quad \forall \mu, \sigma^2
\end{aligned}$$

et $[I^-(\mathbf{X}), I^+(\mathbf{X})] = \left[\bar{X} \pm t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right]$ est un intervalle de confiance pour μ au niveau de confiance $(1-\alpha)$.

IC pour la moyenne d'un échantillon de loi quelconque

Soient $X_1, \ldots, X_n$ i.i.d. où $\sigma^2 = \mathrm{Var}[X_i] < \infty$. Posons $\mu = \mathrm{E}[X_i]$ et $\bar{X}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$ On sait grâce au TCL (et au lemme de Slutzky) que $\frac{\bar{X}^{(n)} - \mu}{\frac{S}{\sqrt{n}}} \to \mathcal{N}(0,1)$. On a donc que

$$\mathrm{P}_\mu\left[z_{\alpha/2} \leq \frac{\bar{X}^{(n)} - \mu}{S/\sqrt{n}} \leq z_{1-\alpha/2}\right] \simeq 1-\alpha \quad \forall \mu$$

$$\mathrm{P}_\mu\left[z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \bar{X}^{(n)} - \mu \leq z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right] \simeq 1-\alpha \quad \forall \mu$$

$$\mathrm{P}_\mu\left[\bar{X}^{(n)} - z_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}^{(n)} - z_{\alpha/2}\frac{S}{\sqrt{n}}\right] \simeq 1-\alpha \quad \forall \mu$$

$$\mathrm{P}_\mu\left[\underbrace{\bar{X}^{(n)} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}}_{=:I^-(\mathbf{X})} \leq \mu \leq \underbrace{\bar{X}^{(n)} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}}_{=:I^+(\mathbf{X})}\right] \simeq 1-\alpha \quad \forall \mu$$

et $[I^-(\mathbf{X}), I^+(\mathbf{X})] = \left[\bar{X} \pm z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$ est un intervalle de confiance pour μ au niveau de confiance $(1-\alpha)$.

Problèmes de test sur une proportion

L'estimateur

$$\hat{p} := \frac{1}{n} \sum X_i$$

d'une proportion p possède de nombreuses propriétés désirables. Sa loi exacte est

$$n\hat{p} \sim \mathrm{Bin}(n, p),$$

et sa loi approchée (approximation généralement considérée satisfaisante pourvu que $np(1-p) > 9$)

$$\hat{p} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

Cette approximation repose sur le théorème central-limite qui nous dit que $n^{1/2}(\hat{p} - p)$ est asymptotiquement $\mathcal{N}(0, p(1-p))$. En tant qu'estimateur de p, $\hat{p}$ est

(i) sans biais

(ii) fortement convergent

(iii) exhaustif et efficace

(iv) solution unique de l'équation de vraisemblance.

Problèmes de test sur une proportion

Tests d'hypothèses (n "grand")

Pour les trois types de test ($H_0 : p \geq p_0$; $H_0 : p \leq p_0$; $H_0 : p = p_0$), la statistique de test est la même, ainsi que la loi utilisée pour le calcul des valeurs critiques.

Statistique de test : $$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Loi sous $p = p_0$:

$$\hat{p} \approx \mathcal{N}\left(p_0, \frac{p_0(1-p_0)}{n}\right) \qquad \text{ou} \qquad Z \approx \mathcal{N}(0,1)$$

- Règle de comportement : dans le problème unilatéral

$$\begin{cases} H_0 & : \quad p \leq p_0 \\ H_1 & : \quad p > p_0, \end{cases}$$

RH_0 au niveau de probabilité α si

- $Z > z_{1-\alpha}$
- la p-valeur $1 - \Phi(Z)$ est inférieure à α
- $\hat{p} > p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$.

Problèmes de test sur une proportion

- Règle de comportement : dans le problème unilatéral (symétrique du précédent)

$$\begin{cases} H_0 & : \quad p \geq p_0 \\ H_1 & : \quad p < p_0, \end{cases}$$

RH_0 au niveau de probabilité α si

  - $Z < -z_{1-\alpha}$
  - la p-valeur $\Phi(Z)$ est inférieure à α
  - $\hat{p} < p_0 - z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$

- Règle de comportement : dans le problème bilatéral

$$\begin{cases} H_0 & : \quad p = p_0 \\ H_1 & : \quad p \neq p_0, \end{cases}$$

RH_0 au niveau de probabilité α si

  - $Z \notin \left[\pm z_{1-\alpha/2}\right]$
  - la p-valeur $2(1 - \Phi(|Z|))$ est inférieure à α
  - $\hat{p} \notin \left[p_0 \pm z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}\right]$.

IC pour une proportion (échantillon de Bernoulli)

Soit un échantillon de Bernoulli $X_1, \ldots, X_n$ iid $\mathrm{Bin}(1, p)$, $p \in (0,\ 1)$. On a que pour $\hat{p} = \bar{X}^{(n)}$ que

$$\mathrm{E}[\bar{X}^{(n)}] = p \qquad\qquad \mathrm{Var}[\bar{X}^{(n)}] = \frac{p(1-p)}{n}$$

Le TCL affirme que $\frac{\bar{X}^{(n)} - p}{\sqrt{\frac{p(1-p)}{n}}} \to \mathcal{N}(0,1)$. En pratique, on pourra utiliser cette approximation asymptotique si $np(1-p) > 9$ ou $n < 20, np < 10$ et $n(1-p) < 10$.

Ainsi, on peut écrire:

$$\mathrm{P}_p \left[z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{1-\alpha/2} \right] \simeq 1 - \alpha \qquad \forall p \in (0,1)$$

$$\mathrm{P}_p \left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right] \simeq 1 - \alpha \qquad \forall p \in (0,1).$$

Il y a un problème car les bornes dépendent du paramètre inconnu p !!!!

IC pour une proportion (échantillon de Bernoulli)

On peut monter que ce paramètre p inconnu peut être remplacé par $\hat{p}$ sans modifier (asymptotiquement) l'intervalle de confiance.

Pour n "grand" ($np(1-p) > 9$), la construction d'un intervalle de confiance peut êre fondée sur la loi normale approchée de $\hat{p}$:

$$\left[\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

(au niveau de confiance (asymptotique ou approchée) $1-\alpha$).

Pour n "petit", cette construction doit être fondée sur la loi binomiale exacte de $n\hat{p}$; les intervalles recherchés s'obtiennent par lecture de tables et d'abaques.

Reprenons nos exemples...

Exercice 1. Supposons que nous prenions un échantillon de 100 Belges. L'âge moyen échantillon est de $\bar{X} = 38,5$ ans et la variance échantillon est $S = 6$. Notons μ l'âge moyen population belge.

(i) Testez l'hypothèse nulle $\mathcal{H}_0 : \mu \leq 40$ contre $\mathcal{H}_1 : \mu > 40$ au niveau $\alpha = .05$.

(ii) Construire un intervalle de confiance pour μ.

Exercice 2. Supposons que nous prenions un échantillon de 100 Français. Parmi ces personnes sélectionnées, une proportion $\hat{p} = 52$ voteront pour le candidat A. Soit p la vraie proportion de Français qui voteront pour ce même candidat A.

(i) Testez l'hypothèse nulle $\mathcal{H}_0 : p \geq .5$ contre $\mathcal{H}_1 : p < .5$ au niveau $\alpha = .05$.

(ii) Construire un intervalle de confiance pour p.

Solutions, discussions.