Eléments de statistiques pour les data sciences

Cours 10: tests et ensembles de confiance

Dr. Ph. Müllhaupt

IGM - EPFL

—

plan

test du rapport de vraisemblance

formulation générale

- ensemble de paramètres Ω, sous ensemble $\omega \subset \Omega$ associè à H_1,
- hypothèses composites $H_0 : \theta \in \omega$ vs. $H_1 : \theta \in \Omega \setminus \omega$
- $\mathrm{x} \in \mathcal{X}$
- rapport de vraisemblance

$$\lambda(\mathrm{x}) = \frac{\sup_\omega L(\theta|\mathrm{x})}{\sup_\Omega L(\theta|\mathrm{x})}$$

- région de rejet (= ensemble critique C)

$$\mathcal{R} = C = \{\mathrm{x}|\lambda(x) \leq k\} \qquad 0 \leq k \leq 1$$

- ensemble d'acceptation

$$\mathcal{A} = \{\mathrm{x}|\lambda(x) > k\} \qquad 0 \leq k \leq 1$$

test du rapport de vraisemblance

seuil d'acceptation k et probabilité α

- hypothèses composites $H_0 : \theta \in \omega \subset \Omega$ vs. $H_1 : \theta = \Omega \setminus \omega$
- la constante $0 \leq k \leq 1$ est choisie de telle sorte que

$$\sup_{\theta \in \omega} P_\theta(\lambda(x) \leq k) = \alpha$$

test composite

$H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

- rapport de vraisemblance

$$\lambda = \frac{L(\theta | H_0; x)}{\max_{\theta \neq \theta_0} L(\theta; x)}$$

- Le dénominateur est $L(\hat{\theta}_{\text{MLE}})$
- $0 < \lambda \leq 1$
- En prenant le logarithme de la vraisemblance, ceci conduit à

$$\log L(\theta_0 | H_0; x) - \log L(\hat{\theta}_{\text{MLE}}); x)$$

- Taille du test : la constante k est choisie telle que

$$\alpha = \sup_{\theta \in \omega} P_\theta\{\lambda(X) \leq k\}$$

avec ω est l'ensemble des paramètres associés à H_0.

Exemple

$X_1, \ldots, X_n$ i.i.d $\sim \mathcal{N}(\mu, \sigma^2)$, μ inconnu

- On observe la réalisation de n variables gaussiennes indépendantes $\mathcal{N}(\mu, \sigma^2)$.
- On ne connait pas μ, mais σ^2 est connu.
- On aimerait un test $H_0 : \theta = \mu$, $H_1 : \theta \neq \mu$. C'est un test composite.

Comme les variables sont indépendantes, la probabilité jointe est le produit des probabilités individuelles.

$$L(\mu) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} k \exp(-(x_i - \mu)^2/(2\sigma^2))$$

Exemple

Maximum de vraisemblance pour Gaussiennes μ et σ incconus

Supposons que la variance ne soit pas connue également. On ne ne connait ni la moyenne, ni la variance. Calculons l'estimateur de maximum de vraisemblance.

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x_i - \mu)^2/(2\sigma^2))$$

avec $\sigma > 0$, $-\infty < \mu < +\infty$.

$$l(\mu, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

les conditions nécessaires pour $\hat{\sigma}$, $\hat{\mu}$ sont

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2} (\sum x_i - n\mu) \triangleq 0 \\ \frac{\partial l}{\partial \sigma} &= \frac{1}{\sigma^3} [\sum (x_i - \mu)^2 - n\sigma^2] \triangleq 0 \end{aligned}$$

et après résolution on trouve les estimateurs de maximum de vraisemblance

$$\begin{aligned} \hat{\mu} &= \frac{\sum x_i}{n} = \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum (x_i - \bar{x})^2 \end{aligned}$$

A première vue, il semblerait que la vraisemblance dépende de toutes les variables individuelles x_i. On va montrer que la vraisemblance dépend uniquement de deux fonctions de celles-ci...

$$\begin{aligned}(x_i-\mu)^2 &= (x_i-\bar{x}+\bar{x}-\mu)^2\\ &= (x_i-\bar{x})^2+(\bar{x}-\mu)^2+2(x_i-\bar{x})(\bar{x}-\mu)\\ \sum(x_i-\mu)^2 &= \sum(x_i-\bar{x})^2+n(\bar{x}-\mu)^2+2\bar{x}-\mu)\sum(x_i-\bar{x})\end{aligned}$$

et comme $\bar{x}\frac{1}{n}\sum x_i$ on a

$$\sum(x_i-\bar{x})=\sum x_i-n\bar{x}=0$$

et donc

$$\sum(x_i-\mu)^2=\sum(x_i-\bar{x})^2+n(\bar{x}-\mu)^2=n\hat{\sigma}^2+n(\hat{\mu}-\mu)^2$$

de telle sorte que la vraisemblance s'écrit sous la forme

$$\boxed{L(\mu,\sigma)=\sigma^{-n}\exp\left(-\frac{n}{2\sigma^2}[\hat{\sigma}^2+(\hat{\mu}-\mu)^2]\right)}$$

Exemple

variance σ^2 connue

Lorsque la variance σ^2 est connue, la vraisemblance devient alors

$$L(\mu) = \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right\}$$

Elle est maximale lorsque $\hat{\mu} = \bar{x}$ car alors $L(\hat{\mu}) = 1$ et

$$\lambda(\mu) = L(\mu) = \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right\}$$

et la statistique de la déviance prend la forme

$$D = -2\log\lambda = n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 \triangleq Z$$

avec

$$Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$$

On peut maintenant déterminer la distribution de D.

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

$$Z \sim \mathcal{N}(0, 1)$$

comme $D = Z^2$

$$D \sim {\chi_{(1)}}^2$$

Exemple

test pour $H_0 : \mu = \mu_0$

On obtient une valeur de D observée, appelée d

$$d = z_{\text{obs}}^2 \qquad z_{\text{obs}} = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$$

Comme $D = Z^2$ on aura le niveau de signification qui dans le cas composite $H_1 : \mu \neq \mu_0$ sera également la probabilité d'erreur de type I

$$\alpha = \text{NS} = P(D \geq d) = P(|Z| \geq |z_{\text{obs}}|)$$

L'hypothèse H_0 est rejetée lorsque

$$\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} > k$$

On a justifié ce qui était dans l'énoncé de l'Exercice 5, Série 10

Exemple 2

variables distribuées exponentiellement

$H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

- $X_1, X_2, \ldots, X_n \sim \mathscr{E}xp(\lambda)$
- La moyenne est θ et donc $\lambda = \frac{1}{\theta}$ et la distribution de X_i est $f_i(x) = \lambda e^{-\lambda x} = \frac{1}{\theta} e^{-\frac{1}{\theta} x}$ avec $x \in [0; +\infty[$

$$\lambda = \frac{\frac{1}{\theta_0^n} e^{-(\sum x_i)/\theta_0}}{\sup_\theta \frac{1}{\theta^n} e^{-(\sum x_i)/\theta}}$$

$$\theta = \frac{\frac{1}{\theta_0^n} e^{-(\sum x_i)/\theta_0}}{\sup_\theta \frac{1}{\theta^n} e^{-(\sum x_i)/\theta}} = \frac{\frac{1}{\theta_0^n} e^{-\sum x_i/\theta_0}}{\frac{1}{(\sum x_i/n)^n} e^{-n}} = \left(\frac{\sum x_i}{n\theta_0}\right)^n e^n e^{-\sum x_i/\theta_0}$$

Le test est rejeté lorsque la condition suivante est vérifiée

$$C = \{x| \left(\frac{\sum x_i}{\theta_0}\right)^n e^{-\sum x_i/\theta_0} < k\}$$

avec k choisi de telle sorte que

$$P_x\{C\} = \alpha$$

estimateur d'intervalle

non nécessairement symétrique

Soit X_1, X_2, ..., X_n des variables aléatoires.

définition

Un estimateur d'intervalle d'un paramètre réel θ est une paire de fonctions $L(x_1, \dots, x_n)$ et $U(x_1, \dots, x_n)$ d'un échantillon x_1, x_2, ..., x_n qui satisfait

$$L(x_1, \dots, x_n) \leq U(x_1, \dots, x_n), \forall x_1, \dots, x_n$$

Lorsque x_1, ..., x_n sont observés alors on peut faire l'inférence

$$\theta \in [L(x_1, \dots, x_n); U(x_1, \dots, x_n)]$$

estimateur d'intervalle

exemple

Soit $X_1, X_2, X_3, X_4 \sim \mathcal{N}(\mu, 1)$

Un estimateur d'intervalle pour μ est, par exemple,

$$[\bar{X} - 1; \bar{X} + 1]$$

$$
\begin{aligned}
P(\mu \in [\bar{X} - 1; \bar{X} + 1]) &= P(\bar{X} - 1 \leq \mu \leq \bar{X} + 1) \\
&= P\left(-2 \leq \frac{\bar{X} - \mu}{\sqrt{1/4}} \leq 2\right) \\
&= P(-2 \leq Z \leq 2) \\
&= 0.9544
\end{aligned}
$$

probabilité de couverture

définition

Comme l'intervalle dépend des échantillons, il peut ou non couvrir le paramètre θ.

définition

probabilité de couverture

$$
\begin{aligned}
& P_\theta(\theta \in [U(x_1, \ldots, x_n); "(x_1, \ldots, x_n)]) \\
= \quad & P_\theta(L(x_1, \ldots, x_n) \leq \theta, U(x_1, \ldots, x_n) \geq \theta)
\end{aligned}
$$

définition

coefficient de confiance

$$\inf_\theta P_\theta(\theta \in [L(x_1, \ldots, x_n); U(x_1, \ldots, x_n)]$$

ATTENTION : c'est l'intervalle qui est aléatoire, pas le paramètre θ ! ! !

estimateur d'intervalle — intervalle de confiance

- estimateur d'intervalle + coefficient de confiance = intervalle de confiance
- On a déjà utilisé ce concept pour $Z \sim \mathcal{N}(\mu, \sigma)$ et pour T de la distribution de Student $\Rightarrow$ intervalles symétriques.
- On peut utiliser de manière équivalente "estimateur d'intervalle" et "intervalle de confiance"

inversion d'un test normal

variables i.i.d. $\mathcal{N}(\mu, \sigma^2)$

- Soit X_1, X_2, ..., X_n des variables aléatoires identiquement distribuées et indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$
- $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- α fixé
- région de rejet $\{x | |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$

En partant de la région d'acceptation de H_0

$$\mathcal{A} = \left\{ x_1, \ldots, x_n \middle| \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}$$

on a

$$\boxed{P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + z_{\alpha/2} \middle| \mu = \mu_0\right) = 1 - \alpha}$$

remarque importante

Cette équation est vraie *quelle que soit la valeur de* μ_0.

inversion d'un test normal

cela conduit donc à ce que la proposition

$$P_\mu\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

soit vraie.

L'intervalle

$$[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}; x + z_{\alpha/2}\sigma/\sqrt{n}]$$

obtenu en *inversant* la région d'acceptation de H_0 d'un test de taille α s'appelle un intervalle de confiance à $1 - \alpha$ probabilité.

inversion d'un test normal

- région de l'espace des échantillons $\mathcal{X}$ d'acceptation du test H_0

$$\mathcal{A}(\mu_0) = \left\{ (x_1, \ldots, x_n) \middle| \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}$$

- intervalle de confiance, ensemble de l'espace des paramètres, ensemble plausible des μ

$$\mathcal{C}(x_1, \ldots, x_n) = \left\{ \mu \middle| \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}$$

Exemple

variables i.i.d. exponentielles

On revient vers l'exemple 2 des variables exponentielles. Le rapport de vraisemblance conduit à

- Pour θ_0 fixé, l'ensemble d'acceptation s'écrit

$$\mathcal{A}(\theta_0) - \left\{ \mathrm{x} \middle| \left(\frac{\sum x_i}{\theta_0} \right)^n \exp\left(-\sum x_i/\theta_0 \right) \geq k \right\}$$

  avec k choisit de telle sorte que $P_{\theta_0}(\mathrm{X} \in \mathcal{A}(\theta_0))$

- Inversion de l'ensemble d'acceptation conduit à l'intervalle de confiance à $1-\alpha$ probabilité

$$\mathcal{C}(\mathrm{x}) = \left\{ \theta \middle| \left(\frac{\sum x_i}{\theta} \right)^n \exp\left(-\sum x_i/\theta \right) \geq k \right\}$$

- Les échantillons x_i apparaissent dans les expressions précédentes que sous la forme $\sum x_i$. Par la théorème de factorisation c'est un statistique suffisante.

$$\mathcal{C}(\sum x_i) = \{\theta | L(\sum x_i) \leq \theta \leq U(\sum x_i)\}$$

- les fonctions L et U sont déterminées bar l'ensemble d'acceptation $\mathcal{A}(\lambda_0)$ de telle sorte qu'il soit associé à une probabilité $1-\alpha$. De plus

$$\left(\frac{\sum x_i}{L(\sum x_i)}\right)^n \exp(-\sum x_i / L(\sum x_i)) = \left(\frac{\sum x_i}{U(\sum x_i)}\right)^n \exp(-\sum x_i / U(\sum x_i))$$

- posons

$$a \triangleq \frac{\sum x_i}{L(\sum x_i)} \qquad b \triangleq \frac{\sum x_i}{U(\sum x_i)}$$

- $$a^n\, e^{-a} = b^n\, e^{-b}$$

Exemple 2

valeur particulière $n = 2$

- cas particulier $n = 2$
- $\alpha = 0.1$
- $$\sum X_i \sim \Gamma(2, \theta) \qquad \sum X_i/\theta \sim \Gamma(2, 1)$$
- $$P_\theta\left(\frac{1}{a}\sum X_i \leq \theta \leq \frac{1}{b}\sum X_i\right) = P\left(b \leq \frac{\sum X_i}{\theta} \leq a\right) = 1 - \alpha$$
- $$\begin{aligned} P\left(b \leq \frac{\sum X_i}{\lambda} \leq a\right) &= \int_b^a t\, e^{-t}\, dt \\ &= e^{-b}(b+1) - e^{-a}(a+1) \end{aligned}$$

- $a = 5.48$ et $b = 0.441$

- on a

$$P_\theta\left(\frac{1}{5.48}\sum_{i=1}^{2} X_i \le \theta \le \frac{1}{0.441}\sum_{i=1}^{2} X_i\right) = 0.90006$$

quantité pivot

définition

quantité pivot

Une variable aléatoire $Q(X_1, \ldots, X_n, \theta)$ est une quantité pivot si la distribution de $Q(X_1, \ldots, X_n, \theta)$ est indépendante des paramètres.

Quel que soit l'ensemble $\mathcal{A}$,

$$P_\theta(Q(X_1, \ldots, X_n, \theta) \in \mathcal{A})$$

ne peut pas dépendre des paramètres.

On cherche le pivot de telle sorte que

$$\{\theta | Q(x_1, \ldots, x_n), \theta) \in \mathcal{A}\}$$

est un ensemble estimateur du paramètre θ.

quantité pivot

exemple $T \sim \Gamma(n, \lambda)$

- $X_1, \ldots, X_n$ i.i.d. $\mathscr{E}xp(\lambda)$
- $T = \sum X_i$ est une statistique suffisante pour λ
- $T \sim \Gamma(n, \lambda)$
- $$Q(T, \lambda) \sim \Gamma(n, \lambda(2/\lambda)) = \Gamma(n, 2)$$

  ne dépend plus de λ et devient une quantié pivot de distribution
- $$\Gamma(n, 2) = \chi^2_{(2n)}$$

quantité pivot

exemple $\mathcal{N}(\mu, \sigma^2)$

- σ^2 inconnu
- le pivot

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

-

$$P\left(-a \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq a\right) = P(-a \leq T_{n-1} \leq a)$$

- $\forall \alpha$,

$$a = t_{n-1,\alpha/2}$$

- l'intervalle de confiance à $1 - \alpha$ probabilité

$$\left\{\mu \middle| \bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right\}$$

quantité pivot

suite de l'exemple

- on aimerait estimer σ^2
- pivot

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

- $$P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = P(a \leq \chi^2_{(n-1)} \leq b) = 1-\alpha$$

- on peut inverser cet ensemble pour obtenir un intervalle de confiance

$$\left\{\sigma \middle| \sqrt{\frac{(n-1)s^2}{b}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{a}}\right\}$$

- un choix possible de a et b

$$a = \chi_{(n-1),1-\alpha/2} \qquad b = \chi_{(n-1),\alpha/2}$$

quantité pivot

retour à l'exemple des distributions exponentielles

- on a obtenu un intervalle de confiance en inversant un test de rapport de vraisemblance $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$
- pivot avec $T = \sum X_i$

$$Q(T, \lambda) = \frac{2T}{\lambda} \sim \chi^2_{(2n)}$$

-

$$P_\lambda(a \leq Q(T, \lambda) \leq b) = P_\lambda\left(a \leq \frac{2T}{\lambda} \leq b\right) = P(a \leq \chi^2_{(2n)} \leq b) = 1 - \alpha$$

- l'inversion de l'ensemble d'acceptation

$$\mathcal{A}(\lambda) = \left\{ t \middle| a \leq \frac{2t}{\lambda} \leq b \right\}$$

- conduit à l'ensemble de confiance

$$\mathcal{C}(t) = \left\{ \lambda \middle| \frac{2t}{b} \leq \chi^2_{(2n)} \leq \frac{2t}{a} \right\} = 1 - \alpha$$