Eléments de statistiques pour les data sciences

Cours 9: tests d'hypothèses - Neyman-Pearson - rapport de vraisemblance

Dr. Ph. Müllhaupt

IGM - EPFL

—

plan

1. les tests de signification
2. la vraisemblance
3. tests statistiques d'une hypothèse versus une alternative
4. l'hypothèse alternative H_1
5. théorie de Neyman-Pearson
6. test statistique du rapport de vraisemblance, déviance
    - test de signification basé sur le rapport de vraisemblance

les tests de signification

ce que l'on a rencontré jusqu'à présent

- une seule hypothèse H_0
- une statistique permettant d'inférer sur l'hypothèse
- une distribution associée à la statistique Z, T, χ^2
- un seuil de signification α (i.e. 5 %)

OBJECTIF :

- Vérifier si les données observées donnent de l'évidence contre l'hypothèse.
- Est-ce que les différences observées sont plutôt le fruit du hasard ou d'une erreur de modèle ?
- Encourager à poursuivre les analyses, si pas assez d'évidence concernant l'hypothèse, en effectuant de nouvelles expériences,
- L'analyse ne conduit pas à une décision dure ; difficulté d'interprétation.

la vraisemblance

pour construire des estimateurs, MLE

- Une expérience conduit à obtenir un échantillon x.
- La structure de probabilité étant définie avant l'expérience, les données conduisent à déterminer les paramètres θ rendant le plus vraisemblable les données observées (l'échantillon x).
- Si on multiplie par une constante positive (ou une fonction positive qui ne dépend pas des paramètres) la probabilité $p(x, \theta)$

$$\hat{\theta}_{\mathrm{MLE}} \triangleq \arg\max_{\theta \in \Theta} k\, p(x, \theta) = \arg\max_{\theta \in \Theta} L(\theta)$$

  est appelé l'estimateur du maximum de vraisemblance.
- Pour des raisons de calcul, on introduit également le logarithme de la vraisemblance.

$$l(\theta) \triangleq \log L(\theta)$$

l'hypothèse alternative H_1

On introduit une hypothèse alternative appelée H_1.

- deux hypothèses en concurrence H_0 et H_1
- Est-ce que l'échantillon est mieux décrit selon H_0 ou H_1 ?
- On peut rejeter H_0 alors qu'elle est vraie, probabilité α, erreur de type I.
- On peut accepter H_0 alors que H_1 est vraie (H_0 est fausse), probabilité β, erreur de type II.

	accepter H_0	accepter H_1
H_0 est vraie	Décision correcte	Erreur I
H_1 est vraie	Erreur II	Décision correcte

une asymétrie du problème est introduite

H_0 est plus importante que H_1

Une asymétrie est introduite :

- On aimerait contrôler α en fixant α.
- On minimise β (maximise la puissance $1-\beta$) sous contrainte de α constant.
- Cela aboutit $\Rightarrow$ la théorie de Neyman-Pearson

partition de l'espace des paramètres

définition équivalente de H_0 et H_1

Pour aboutir à une formulation générale de la théorie de Neyman-Pearson, on va reformuler les hypothèses H_0 et H_1 comme une partitition de l'espace des paramètres.

- Une hypothèse nulle H_0 est dite simple si elle invoque uniquement la valeur précise d'un seul paramètre.
- L'hypothèse alternative H_1 est dite simple si elle invoque uniquement la valeur précise d'un seul paramètre.
- Tous les autres cas sont dits composites.

de manière générale

- H_0 est équivalent à dire que $\theta \in \omega \subset \Theta$
- H_1 est équivalent à dire que $\theta \in \Theta \setminus \omega$

H_0 est plus important que H_1

on fixe α

- L'hypothèse H_0 est plus importante que H_1.
- On aimerait donc contrôler α (appelé la taille du test)
- On aimerait également minimiser β.
- En conséquence : il faut maximiser $1 - \beta$, appelé la puissance du test.

La région critique C (espace des échantillons)

rejet de H_0 alors que H_0 est vraie

la région critique C

On désigne par C la région de l'espace des échantillons $C \subset \mathcal{X}$ qui est associée au rejet de H_0 alors que H_0 est vraie.

Notations :

- L'espace des paramètres est Θ. ω est associée à l'hypothèse H_0 et $\Theta \setminus \omega$ représente l'hypothèse H_1.
- Lorsque H_0 est vraie, et qu'il s'agit d'une hypothèse simple $\theta = \theta_0$, on notera la probabilité sous ce modèle comme p_{θ_0} ou lorsqu'il n'y a pas de confusion p_0.
- Lorsque H_1 est vraie, et qu'il s'agit d'une hypothèse simple $\theta = \theta_1$, on notera la probabilité sous ce modèle comme p_{θ_1} ou lorsqu'il n'y a pas de confusion p_1.

La région critique C (espace des échantillons)

rejet de H_0 alors que H_0 est vraie

Une partition de l'espace des échantillons $\mathcal{X}$ conduit à étiqueter les échantillons selon si on décide H_0 ou si on décide H_1, indépendamment si H_0 est vraie ou si H_1 est vraie. La partitition est donnée par une statistique.

- La région critique C correspond à l'étiquetage H_1 de l'espace des échantillons. (C'est un sous-ensemble de tous les échantillons possibles.)
- La probabilité d'erreur de type I s'ecrit :

$$\alpha \triangleq P_{\theta_0}(C)$$

Le lemme de Neymann-Pearson

intuition

- Les hypothèses H_0 et H_1 sons simples.
- L'idée est que plus $p_{\theta_1}/p_{\theta_0}$ est grand plus l'hypothèse H_1 semble plausible par rapport à H_0.
- Toute la subtilité est de garantir une condition pour que la probabilité d'erreur de type I, à savoir α, soit sous contrôle.
- On décidera sur H_1 lorsque la statistique utilisée conduira à ce que $p_{\theta_1}/p_{\theta_0}$ dépassera un seuil critique k.

Le lemme de Neyman-Pearson

lemme de Neyman-Pearson

Si les hypothèses suivantes sont respectées :

- Soit C une région de l'espace des échantillon $\mathcal{X}$ telle que $P_{\theta_0}(C) \leq \alpha$.
- Supposons qu'il existe une région C^* de $\mathcal{X}$ de la forme $C^* = \{x | p_{\theta_1}(x)/p_{\theta_0}(x) \geq k\}$ telle que $P_{\theta_0}(C^*) = \alpha$

alors

$$P_{\theta_1}(C^*) \geq P_{\theta_1}(C)$$

Explication : il n'est pas possible d'obtenir plus de puissance du test par rapport à celle donnée par la région critique C^*. Cette région donne la probabilité β la plus faible en garantissant une probabilité α donnée (erreur de type I sous contrôle au niveau maximum α).

Le lemme de Neyman-Pearson

démonstration

démonsitration

On suppose que C^* existe. $\bar{C}$ désigne le complémentaire de C et $\bar{C}^*$ désigne le complémentaire de C^*.

$$P_{\theta_1}(C^*) - P_{\theta_1}(C) = \int_{C^*\cap\bar{C}} p_1(x)dx - \int_{\bar{C}^*\cap C} p_1(x)dx$$

Dans $C^* \cap \bar{C}$ on a la propriété $p_1(x) \geq k\, p_0(x)$ et donc

$$\int_{C^*\cap\bar{C}} p_1(x)dx \geq k \int_{C^*\cap\bar{C}} p_0(x)dx$$

De manière similaire

$$\int_{\bar{C}^*\cap C} p_1(x)dx < k \int_{\bar{C}^*\cap C} p_0(x)dx$$

en conséquence

$$
\begin{aligned}
P_{\theta_1}(C^*) - P_{\theta_1}(C) &\geq k\left[\int_{C^*\cap\bar{C}} p_0(x)dx - \int_{\bar{C}^*\cap C} p_0(x)dx\right] \\
&= k\left[\int_{C^*} p_0(x)dx - \int_{C} p_0(x)dx\right] \\
&= k[P_{\theta_0}(C^*) - P_{\theta_0}(C)] \\
&\geq 0
\end{aligned}
$$

étant donné que $k \geq 0$, $P_{\theta_0}(C^*) = \alpha$ et $P_{\theta_0}(C) \leq \alpha$. On arrive donc à la conclusion du lemme

$$
P_{\theta_1}(C^*) \geq P_{\theta_1}(C) \qquad \text{C.Q.F.D.}
$$

Le lemme de Neyman-Pearson

commentaire

Le lemme indique une méthode pour trouver la constante k :

- Choisir k de telle sorte que

$$P_{\theta_0}\left\{x \left| \frac{p_1(x)}{p_0(x)} \geq k \right.\right\} = \alpha$$

- Dans beaucoup de cas $p_1(x)/p_0(x)$ est une fonction croissante, on commence donc par choisir k petit et l'on augmente jusqu'à atteindre la valeur α désirée.
- La condition $p_1(x)/p_0(x) \geq k$ est ensuite montrée être équivalente à une condition sur une statistique $T(x) \geq k'$ avec k' une nouvelle constante fonction de k. On dimensionne alors k' pour que α soit attient. Lorsque la distribution de T est connue (par exemple Student t, gaussienne Z, ou Pearson χ^2) on utilise les tables correspondantes.

Exemple 1

application du lemme de Neyman-Pearson

- 3 variables aléatoires indépendantes $X_i \sim \mathscr{B}er(p)$, $i = 1, 2, 3$
- observation : $X = \sum_{i=1}^{3} X_i \sim \mathscr{B}in(n, p)$, $x = 0, 1, 2, 3$
- $\theta \triangleq p$
- $H_0 : \theta_0 = \frac{1}{4}$ $\qquad$ $H_1 : \theta_1 = \frac{3}{4}$
- $\alpha = 0.05$

avec $x = 0, 1, 2, 3$, on a

$$
\begin{aligned}
\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} &= \frac{\theta_1^x (1-\theta_1)^{n-x}}{\theta_0^x (1-\theta_0)^{n-x}} = \frac{\left(\frac{3}{4}\right)^x \left(1-\frac{3}{4}\right)^{3-x}}{\left(\frac{1}{4}\right)^x \left(1-\frac{1}{4}\right)^{3-x}} \\
&= \frac{\left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{3-x}}{\left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{3-x}} = \frac{\left(\frac{1}{4}\right)^{x+3-x}}{\left(\frac{1}{4}\right)^{x+3-x}} \frac{3^x 1^{3-x}}{1^x 3^{3-x}} = 3^{2x-3} 1^{1-2x} \\
&= 3^{2x-3}
\end{aligned}
$$

On constate que c'est une fonction croissante de x :

$$
\begin{aligned}
x = 0 \quad &: \quad p_1/p_0 = \frac{1}{27} \\
x = 1 \quad &: \quad p_1/p_0 = \frac{1}{3} \\
x = 2 \quad &: \quad p_1/p_0 = 3 \\
x = 3 \quad &: \quad p_1/p_0 = 27
\end{aligned}
$$

On doit choisir k si possible pour que

$$P_{\theta_0}\left\{x \left| \frac{p_1(x)}{p_0(x)} \geq k \right.\right\} = 0.05$$

Cette condition est équivalente (à cause de la croissance monotone de p_1/p_0) à déterminer une nouvelle constante k' telle que

$$P_{\theta_0}\left\{x | x > k'\right\} = 0.05$$

On examine cas par cas :

- si $k' \in]2, 3[$ alors $\{x|x > k'\} = 3$, ce qui donne $P_{\theta_0}(3) = \left(\frac{1}{4}\right)^3 = \frac{1}{64}$
- si $k' \in]1; 2[$ alors $P_{\theta_0}\{x|x \geq k'\} = P\{x = 2 \text{ ou } x = 3\}$

On calcule successivement :

$$
\begin{aligned}
P_{\theta_0}(x = 2) &= C_2^3 \left(\frac{1}{4}\right)^2 \frac{3}{4} = \frac{9}{64} = 0.140625 \\
P_{\theta_0}(x = 3) &= \frac{1}{64} = 0.015625 \\
P_{\theta_0}(x = 2 \text{ ou } x = 3) &= P_{\theta_0}(x = 2) + P_{\theta_0}(x = 3) \\
&= \frac{9}{64} + \frac{1}{64} = \frac{10}{64} = 0.15625
\end{aligned}
$$

Dans cet exemple, on ne peut pas atteindre précisément $\alpha = 0.05$ car

$$0.015625 < 0.05 < 0.15635$$

Exemple 2

application du lemme de Neyman-Pearson

- soit $X = (X_1, X_2, \ldots, X_n)$ un échantillon avec $X_i \sim \mathcal{N}(\mu, 1)$
- $H_0 : \mu = \theta_0 \qquad H_1 : \mu = \theta_1 \qquad \theta_1 > \theta_0$

$$
\begin{aligned}
\frac{p_1(x)}{p_0(x)} &= e^{-\frac{1}{2}\sum(x_i-\theta_1)^2+\frac{1}{2}\sum(x_i-\theta_0)^2} \\
&\geq k \\
&\Updownarrow \\
n\bar{x}(\theta_1-\theta_0)-\frac{1}{n}(\theta_1^2-\theta_0^2) &\geq k' \\
&\Updownarrow \\
\bar{x} &\geq k''
\end{aligned}
$$

On peut ainsi déterminer un k tel que $P_{\theta_0}\{x|p_1(x)/p_0(x)\geq k\}=\alpha$ si on peut déterminer un k'' tel que $P_{\theta_0}\{x|\bar{x}\geq k''\}=\alpha$

Si θ_0 est la vrai paramètre

$$\bar{x}\sim\mathcal{N}(\theta_0,1/n)\Leftrightarrow\sqrt{n}(\bar{x}-\theta_0)\sim\mathcal{N}(0,1)$$

Soit k_α la valeur de la probabilité de 100α pourcents supérieurs d'une loi gaussienne

$$\begin{aligned} P_{\theta_0}(\sqrt{n}(\bar{x}-\theta_0) \geq k_\alpha) &= \alpha \\ P_{\theta_0}\left\{\bar{x} \geq \frac{k_\alpha}{\sqrt{n}} + \theta_0\right\} &= \alpha \end{aligned}$$

$$C^* = \left\{ x \middle| \bar{x} \geq \frac{k_\alpha}{\sqrt{n}} + \theta_0 \right\}$$

est la région critique du test le plus puissant pour un test au niveau α de signification de l'hyphothèse simple $\mu = \theta_0$ contre l'hypothèse simple $\mu = \theta_1$ avec $\theta_1 > \theta_0$.

Hypothèse simple contre hypothèse composite

Dans le cas d'une hypothèse simple $H_0 : \theta = \theta_0$ et une hypothèse compositie $H_1 : \theta \neq \theta_0$, le lemme de Neyman-Pearson n'indique plus nécessairement le test le plus puissant.

Si on examine le ratio dans le test de Neyman-Pearson il fait apparaître un rapport de vraisemblance.

On va procéder en utilisant un rapport de vraisemblance, mais avec une fraction inverse. Ceci est dû à la connection avec le maximum de vraisemblance.

Test de signification pur et rapport de vraisemblance

Revenons en arrière et considérons un test de signification.

Définissons

$$\lambda = \frac{L(\theta|H_0;x)}{\max_\theta L(\theta;x)} = \frac{L(\theta|H_0;x)}{L(\hat{\theta}_{\text{MLE}};x)}$$

Il s'agit du rapport entre la vraisemblance sous l'hypothèse $\theta = \theta_0$ divisé par le maximum de vraisemblance qui conduit au paramètre $\theta = \hat{\theta}_{\text{MLE}}$.

- Le maximum est pris sur tout l'ensemble des paramètres. Le maximum n'est pas contraint à un sous-ensemble particulier de Θ. On trouve donc $\hat{\theta}_{\mathsf{MLE}}$ rendant maximum la vraisemblance.

- $1 \leq \lambda \leq 0$

- Lorsque H_0 est le plus vraisemblant on aura tendance á ce que $\lambda \to 1$ (à cause de faire apparaîtrea H_0 au numérateur plutôt qu'au dénominateur dans la cas du lemme de Neyman-Pearson).

Définition de la déviance

Cette définition est conséquence du rapport de vraisemblance λ en prenant les logarithmes.

Comme $\lambda < 1$, il est mieux de changer le signe du logarithme pour faire apparaître une quantité positive.

A cause de la relation avec le χ^2 on introduit un facteur 2.

définition de la déviance

$$\boxed{D = -2[l_0 - \hat{l}]} = 2[\hat{l} - l_0] = 2[l(\hat{\theta}) - l(\theta_0)] = -2r(\hat{\theta}_0)$$

avec $l_0 = \log(L(\theta_0; x)$ et $\hat{l} = \log(L(\hat{\theta}_{\mathrm{MLE}}; x)$

Propriété de la déviance

niveau de signification et $D \approx \chi^2_{(1)}$

niveau de signification

$$NS = P\{D \geq D_{\text{obs}} | H_0 \text{ est vraie}\}$$

relation avec le χ^2

Le théorème central limite permet d'établir la propriété

$$\lim_{n\to\infty} P(D \leq d | \theta = \theta_0) = P\{\chi^2_{(1)} \leq d\}$$