Compte-rendu en *Principes et méthodes statistiques*

Analyse de signaux oculométriques

Aurélien PEPIN, Léo DESBUREAUX, Julien LABOURÉ (Ensimag)

5 mai 2017

1 Analyse d'échantillons de loi binomiale négative

QUESTION 1. On suppose dans un premier temps r connu. L'estimateur obtenu par la méthode des moments est noté $\tilde{p}_n$.

> Comme on suppose l'échantillon $X_1, X2, ..., X_n$ indépendantes et de même loi binomiale négative $\mathcal{BN}(r,p)$, on a $E[X] = \frac{r}{p}$. Alors l'estimateur des moments est :
>
> $$\tilde{p}_n = \frac{r}{\overline{X}_n}$$
>
> où $\overline{X}_n$ désigne la moyenne empirique de l'échantillon.

Une deuxième méthode d'estimation ponctuelle est l'estimation par maximum de vraisemblance. On note maintenant l'estimateur trouvé $\hat{p}_n$.

> $$\mathcal{L}(p; x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n; p) = \prod_{i=1}^{n} P(X = x_i; p)$$
>
> Plutôt que de dériver directement ce produit, on préfère maximiser le logarithme de la fonction de vraisemblance $\mathcal{L}$, c'est la **log-vraisemblance** :
>
> $$\begin{aligned} \ln \mathcal{L}(p; x_1, ..., x_n) &= \ln \prod_{i=1}^{n} P(X = x_i; p) = \sum_{i=1}^{n} \ln P(X = x_i; p) \\ &= \sum_{i=1}^{n} \ln \left(\binom{x_i - 1}{r - 1} (1-p)^{x_i - r} p^r \right) \end{aligned}$$
>
> On cherche désormais à obtenir $\hat{p}_n$, valeur qui maximise cette log-vraisemblance. On dérive pour cela l'expression précédente :
>
> $$\begin{aligned} \frac{\partial}{\partial p} \ln \mathcal{L}(p; x_1, ..., x_n) = \frac{\partial}{\partial p} \left(\sum_{i=1}^{n} \ln \binom{x_i - 1}{r - 1} \right) &+ \sum_{i=1}^{n} (x_i - r) \frac{\partial}{\partial p} (\ln(1-p)) + \sum_{i=1}^{n} r \frac{\partial}{\partial p} (\ln p) \\ = 0 \qquad\qquad &- \sum_{i=1}^{n} (\frac{x_i - r}{1-p}) \qquad\qquad + \sum_{i=1}^{n} \frac{r}{p} \end{aligned}$$

Cette expression s'annule sous les conditions suivantes :

$$
\begin{aligned}
-\sum_{i=1}^{n}(\frac{x_i - r}{1-p}) + \sum_{i=1}^{n}\frac{r}{p} = 0 &\iff \sum_{i=1}^{n}\frac{-px_i + pr + (1-p)r}{p(1-p)} = 0 \\
&\iff \sum_{i=1}^{n} -px_i + r = 0 \\
&\iff \sum_{i=1}^{n}\frac{r}{p} = \sum_{i=1}^{n} x_i \\
&\iff n\frac{r}{p} = \sum_{i=1}^{n} x_i \\
&\iff \frac{r}{p} = \overline{X}_n
\end{aligned}
$$

Finalement, on retrouve donc le résultat précédent :

$$\hat{p}_n = \frac{r}{\overline{X}_n} = \tilde{p}_n$$

Les cas aux limites où $p = 0$ ou $p = 1$ correspondent à des situations triviales où tous les X_i sont identiques, il n'y a aucune part d'aléatoire.

QUESTION 2. Pour une suite de variables aléatoires iid $\{X_n\}_{n\geq 1}$, le théorème central-limite exprime la convergence suivante :

$$Z_n = \sqrt(n)\frac{\overline{X}_n - E[X]}{\sigma(x)} \rightarrow \mathcal{N}(0,1)$$

La suite $X_1, ..., X_n$ satisfait les conditions du théorème et on a les données suivantes :

- $E[X] = \frac{r}{p}$
- $\sigma(X) = \frac{\sqrt{(r(1-p))}}{p}$

Dans notre cas :

$$
\begin{aligned}
Z_n &= \sqrt(n)p\frac{\overline{X}_n - \frac{r}{p}}{\sqrt{(r(1-p))}} \\
&= \frac{\sqrt{nr}}{\sqrt{1-p}}(\frac{p}{\hat{p}_n} - 1) \rightarrow \mathcal{N}(0,1)
\end{aligned}
$$

Sachant que $P(\mid Z_n \mid > u_\alpha) = \alpha$, on a :

$$
\begin{aligned}
\frac{\sqrt{nr}}{\sqrt{1-p}} \mid \frac{p}{\hat{p}_n} - 1 \mid > u_\alpha &\iff \sqrt{rn} \mid p - \hat{p}_n \mid > u_\alpha \hat{p}_n \sqrt{1-p} \\
&\iff rn(p-\hat{p}_n)^2 > u_\alpha^2 \hat{p}_n^2 (1-p) \\
&\iff rnp^2 + (-2rn\hat{p}_n + u_\alpha^2 \hat{p}_n^2)p + \hat{p}_n^2(rn - u_\alpha^2) > 0 \\
&\iff p^2 + \hat{p}_n(-2 + \frac{u_\alpha^2 \hat{p}_n}{rn})p + \hat{p}_n^2(1 - \frac{u_\alpha^2 \hat{p}_n}{rn}) > 0
\end{aligned}
$$

Posons $\lambda = \frac{u_\alpha^2 \hat{p}_n}{rn}$, alors on a :

$$
p^2 + p\hat{p}_n(-2 + \lambda \hat{p}_n) + \hat{p}_n^2(1-\lambda) > 0
$$

On obtient ainsi un polynôme en p de degré 2, dont on déduit le discriminant :

$$
\Delta = (\hat{p}_n^2 \lambda)^2 + 1 + \frac{4(1-\hat{p}_n)}{\lambda \hat{p}_n}
$$

Á partir du discriminant, on peut calculer les racines du polynôme qui sont les bornes de l'intervalle de confiance qu'on cherche à déterminer :

$$
IC = \left[\hat{p}_n - \frac{1}{2}\lambda \hat{p}_n^2 \left(1 + \sqrt{1 + \frac{4(1-\hat{p}_n)}{\lambda \hat{p}_n}}\right) ; \hat{p}_n - \frac{1}{2}\lambda \hat{p}_n^2 \left(1 - \sqrt{1 + \frac{4(1-\hat{p}_n)}{\lambda \hat{p}_n}}\right)\right]
$$

QUESTION 3. Pour tracer le graphe de probabilités de la loi géométrique (qui correspond au cas $r = 1$), on cherche en premier lieu des fonctions h, α, g, β telles que :

$$
h[F(k)] = \alpha(p)\ g(k) + \beta(p)
$$

Se référer à : `P1_Q3_Graphe_Probabilites.r`

La fonction de répartition de la loi géométrique est :

$$
\begin{aligned}
F_{\mathcal{G}}(k) = 1 - (1-p)^k &\iff 1 - F_{\mathcal{G}}(k) = (1-p)^k \\
&\iff \ln(1 - F_{\mathcal{G}}(k)) = k \ln(1-p)
\end{aligned}
$$

Par identification, on établit les correspondances suivantes :

— $h[F_{\mathcal{G}}(k)] = \ln(1 - F_{\mathcal{G}}(k))$
— $\alpha(p) = \ln(1-p)$
— $g(k) = k$
— $\beta(p) = 0$

Le graphe de probabilités de $F_{\mathcal{G}}(k)$ est donc le nuage de points :

$$
(g(k_i^*); h(\frac{i}{n}) = (k_i^*; \ln(1 - \frac{i}{n}))\ \forall i \in [\![1; n-1]\!]
$$

INSERER ICI DE MIRIFIQUES GRAPHIQUES LE GROUPE 1 EST MEILLEUR QUE LE GROUPE 2
ON PREND ENSUITE LA PENTE DU GROUPE 1 ET ON L'INJECTE EN PARAMETRE D'UNE SIMU DE 10 000 DONNEES
QUESTION 4. Sous-partie a).

$$\begin{aligned}
\frac{P(X=x)}{P(X=x+1)} &= \frac{\binom{x-1}{r-1}(1-p)^{x-r}p^r}{\binom{x}{r-1}(1-p)^{x+1-r}p^r} \\
&= \frac{\frac{(x-1)!}{(r-1)!\ (x-r)!}(1-p)^{x-r}p^r}{\frac{x!}{(r-1)!\ (x-r+1)!}(1-p)^{x+1-r}p^r} \\
&= \frac{x-r+1}{x(1-p)}
\end{aligned}$$

Sous-partie b). On définit $g(x) = x\frac{P(X=x)}{P(X=x+1)} = \frac{1}{1-p}x + \frac{1-r}{1-p}$.

> On affiche le nuage de points $(x, g(x))$. Á partir d'une régression linéaire, on trace la droite au sens des moindres carrés $y = ax + b$ pour obtenir un coefficient directeur :
>
> $$a = \frac{1}{1-p} \iff p = 1 - \frac{1}{a}$$

Sous-partie c). COMMENCER A r!!! (via paramètre de la loi)
QUESTION 5. On suppose désormais que le paramètre r est inconnu. On va donc chercher à estimer simultanément les deux paramètres (r, p) via la méthode des moments.

Pour cela, on rappelle les égalités suivantes $\begin{cases} E[X] = \frac{r}{p} \\ Var(X) = \frac{r(1-p)}{p^2} \end{cases}$

> $$\begin{aligned}
> \frac{E[X]}{Var(X)} = \frac{p}{1-p} &\iff E[X](1-p) = Var(X)p \\
> &\iff E[X] = p(E[X] + Var(x))
> \end{aligned}$$
>
> On en déduit ainsi que :
>
> $$\tilde{p}_n = \frac{\overline{X}_n}{\overline{X}_n + S_n^2}$$
>
> $$\tilde{r}_n = \frac{\overline{X}_n^2}{\overline{X}_n + S_n^2}$$

QUESTION 6. ETAPE 1. On détermine $z = \min_i x_i$. En effet, la valeur de r ne peut pas être supérieure à la valeur des x_i.

ETAPE 2. Pour chaque valeur de r possible, donc pour tout $i \in [\![1; z]\!]$, on calcule l'estimateur $\hat{p}_{n,i} = \frac{i}{\overline{X}_n}$ (sur le modèle de la question 1).

ETAPE 3. Pour chaque couple $(r_i, p_i) = (i, \hat{p}_{n,i})$, on calcule sa vraisemblance.

ETAPE 4. On en déduit $(\hat{r}_n, \hat{p}_n) = \mathcal{L}(r_i, p_i; x_1, ..., x_n)$. faire une petite phrase de conclusion

2 PARTIE 2 ECRIRE LE TITRE

QUESTION 1. Faire comme pour l'exercice sur la pollution. **QUESTION 2**. à voir en fonction de la q4.