Entrega 1 Estadística

Guillem Tutusaus i Alcaraz, 1533701

03/04/2024

1. Sigui $x = (x_1, \ldots, x_n)$, $n > 2$ observacions independents, idènticament distribuïdes amb funció de probabilitat

$$\mathbb{P}(X_i = x_i) = p(1-p)^{x_i - 1}$$

on $x_i \in \mathbb{N}$, $p \in (0, 1)$.

(a) Calculem el valor esperat de $S_n(X) := \sum_{i=1}^{n} X_i$.

$$\mathbb{E}(S_n(X)) = \mathbb{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n\mathbb{E}(X_1)$$

on en aquesta darrera igualtat hem fet servir que $X_1, \ldots, X_n$ son idènticament distribuïdes. Ara,

$$\mathbb{E}(X_1) = \sum_{i \geq 1} ip(1-p)^{i-1} = p\frac{1}{p^2}$$

per tant, $\mathbb{E}(X_1) = \frac{1}{p}$. D'aquesta manera,

$$\mathbb{E}(S_n(X)) = \frac{n}{p}$$

També podriem haver observat que $X_1 \sim Geom(p)$ que sabem que si aquesta té suport $\mathbb{N}$, aleshores $\mathbb{E}(X_1) = \frac{1}{p}$.

(b) Trobem un estadístic suficient per p.
Un estadístic T és suficient per un paràmetre θ si la probabilitat condicionada de la nostra mostra, donat T no depèn del paràmetre. En el nostre cas, el paràmetre és p. Escrivint la probabilitat condicionada tenim

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | T = t) = \\ \frac{\mathbb{P}(T = t | X_1 = x_1, \ldots, X_n = x_n)\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}{\mathbb{P}(T = t)}$$

pel teorema de Bayes.

Notem que, al ser les X_i's independent idènticament distribuïdes podem escriure el següent

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) = \prod_{i=1}^{n} p(1-p)^{x_i - 1}$$

A més, si $X_i \sim Geom(p)$ per a tot i, aleshores considerant $S_n(X) := \sum_{i=1}^n X_i$ és fàcil veure per inducció sobre n que $S_n \sim NB(n,p)$. D'aquesta manera, considerant l'estadístic $S_n(X) := \sum_{i=1}^n X_i$ tenim

$$\mathbb{P}(S_n = t) = \binom{t-1}{n-1} p^n (1-p)^{t-n}$$

on $t \in \{n, n+1, \ldots\}$. Per últim, observem que

$$\mathbb{P}(S_n = t | X_1 = x_1, \ldots, X_n = x_n) = \begin{cases} 1 & \text{si } t = \sum_{i=1}^n x_i \\ 0 & \text{altrament} \end{cases}$$

Per tant,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | S_n = t) = \frac{\prod_{i=1}^n p(1-p)^{x_i - 1}}{\binom{t-1}{n-1} p^n (1-p)^{t-n}}$$

$$= \frac{p^n (1-p)^{\sum_{i=1}^n x_i - n}}{\binom{t-1}{n-1} p^n (1-p)^{t-n}} = \frac{(1-p)^{\sum_{i=1}^n x_i}}{\binom{t-1}{n-1} (1-p)^t} = \frac{1}{\binom{t-1}{n-1}}$$

com estem suposant que $t = \sum_{i=1}^n x_i$ obtenim el resultat $\frac{1}{\binom{t-1}{n-1}}$ que no depèn de p i per tant això demostra que $S_n(X)$ és un estadístic suficient per p.

(c) Trobem un estimador no esbiaixat per $\frac{1}{p}$ en termes de $S_n(X)$.

Per (1a) sabem que $\mathbb{E}(S_n(X)) = \frac{n}{p}$, per tant considerant

$$T_n(X) := \frac{1}{n} S_n(X)$$

ja ho hauriem de tenir. Comprovem-ho,

$$\mathbb{E}(T_n(X)) = \frac{1}{n} \mathbb{E}(S_n(X)) = \frac{1}{n} \frac{n}{p} = \frac{1}{p}$$

com volíem veure. En particular, l'estimador buscat és

$$T_n(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(d) Trobem un estimador no esbiaixat per p en termes de $S_n(X)$.

Recordem que si X és una variable aleatòria discreta i $h : \mathbb{R} \to \mathbb{R}$ una funció tal que

$$\sum_{k \in S} |h(k)| \mathbb{P}(X = k) < \infty$$

aleshores definim l'esperança de $h(X)$ com

$$\mathbb{E}(h(X)) = \sum_{k \in S} h(k) \mathbb{P}(X = k)$$

Ara, tenint present la indicació donada considerem l'estadístic

$$Q_n(X) := \frac{1}{S_n(X) - 1}$$

i en calculem la seva esperança. Per això cal recordar que $S_n \sim NB(n, p)$ té funció de probabilitat $\mathbb{P}(S_n = k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}$, $k \in \{n, n+1, \dots\}$. D'aquesta manera,

$$\begin{aligned}
\mathbb{E}(Q_n) &= \mathbb{E}\Big(\frac{1}{S_n - 1}\Big) \\
&= \sum_{k \geq n} \frac{1}{k-1} \mathbb{P}(S_n = k) \\
&= \sum_{k \geq n} \frac{1}{k-1} \binom{k-1}{n-1} p^n (1-p)^{k-n} \\
&= \frac{1}{(n-1)!} \Big(\frac{p}{1-p}\Big)^n \sum_{k \geq n} \frac{(k-1)!}{(k-1)(k-n)!} (1-p)^k \\
&= \frac{1}{(n-1)!} p^n \sum_{i \geq 0} \frac{(i+n-2)!}{i!} (1-p)^i \\
&= \frac{1}{(n-1)!} p^n \frac{(n-2)!}{p^{n-1}} \\
&= \frac{p}{n-1}
\end{aligned}$$

D'aquesta manera i en virtud de les observacions fetes anteriorment, tenim que l'estimador

$$(n-1) Q_n(x)$$

verifica el que volem. Per completitud calculem la sèrie

$$\sum_{i \geq 0} \frac{(i+n-2)!}{i!} (1-p)^i$$

Comencem reescrivint el numerador com

$$\frac{(i+n-2)!}{i!} = (i+1)(i+2)\dots(i+n-2)$$

Agafant ara la sèrie de potències $f(x) = \sum_{i\geq 0} x^i = \frac{1}{1-x}$ per $|x| < 1$ i prenent-ne la $(n-2)$-derivada veiem

$$f^{(n-2)}(x) = \sum_{i\geq n-2} i(i-1)\dots(i-n+3)x^{i-(n-2)} = \frac{(n-2)!}{(1-x)^{n-1}}$$

Així doncs la nostra sèrie és

$$\sum_{i\geq 0} \frac{(i+n-2)!}{i!}(1-p)^i = f^{(n-2)}(1-p) = \frac{(n-2)!}{p^{n-1}}$$

2. Una mostra $X_1, \dots, X_n$ s'extreu d'una distribució normal $N(\theta, \theta^2)$.

   (a) Troba un interval de confiança del 90% per la mitjana poblacional θ.

   Per calcular un interval de confiança del 90% per θ considerem el següent estimador.

$$\frac{\bar{X}-\theta}{\frac{S_X}{\sqrt{n}}}$$

   Sabem de teoria que aquest estimador té distribució t-Student amb $n-1$ graus de llibertat, i.e.

$$\frac{\bar{X}-\theta}{\frac{S_X}{\sqrt{n}}} \sim t^{n-1}$$

   Això ens permet construïr un interval de confiança per θ. Un interval de confiança del 90% ($\alpha = 0.1$) per a la distribució t^{n-1} és

$$(t_{0.05}^{n-1}, t_{0.95}^{n-1}) = (-t_{0.95}^{n-1}, t_{0.95}^{n-1})$$

   on $t_{0.95}^{n-1}$ és el quantil 95% de la distribució amb $n-1$ graus de llibertat. D'aquesta manera, podem construïr un interval per θ

$$\begin{aligned} -t_{0.95}^{n-1} &< \frac{\bar{X}-\theta}{\frac{S_X}{\sqrt{n}}} < t_{0.95}^{n-1} \\ -t_{0.95}^{n-1}\frac{S_X}{\sqrt{n}} &< \bar{X}-\theta < t_{0.95}^{n-1}\frac{S_X}{\sqrt{n}} \\ \bar{X}-t_{0.95}^{n-1}\frac{S_X}{\sqrt{n}} &< \theta < \bar{X}+t_{0.95}^{n-1}\frac{S_X}{\sqrt{n}} \end{aligned}$$

(b) Volem fer un test per determinar si existeix un canvi respecte la mitjana poblacional i un conjunt de dades anteriorment obtinguda. Quina mida de la mostra necessitem per detectar un canvi de 0.1 amb una significació del 95% i una potència del 90% si sabem que $\bar{X} = 12$?

Per tal de respondre a aquesta pregunta ens plantejem el següent test per a θ amb variància desconeguda.

$$\begin{cases} H_0 : \theta - \theta_0 = 0 \\ H_1 : |\theta - \theta_0| \geq 0.1 \end{cases}$$

On θ_0 representa la mitjana poblacional del primer test realitzat. Notem que al disposar de tan poques hipòtesis, al fer les dues observacions, obtindrem $\hat{X}, \hat{X}_0, S_X^2$ i $S_{X_0}^2$.

Ara, sabem pel Seminari 3 que sota les condicions anteriors

$$\frac{(\hat{X} - \hat{X}_0) - (\theta - \theta_0)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_{X_0}^2}{n_{X_0}}}} \sim t^k$$

on no estem suposant que coneguem les variancies de les dues mitjanes ni que siguin la mateixa ni tampoc que les mides de les mostres siguin iguals (per això n_X i n_{X_0}). Aquest estadístic segueix una llei t^k amb k graus de llibertat. Per trobar els graus de llibertat en aquest cas cal resoldre:

$$k = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_{X_0}^2}{n_{X_0}}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_{X_0}^2}{n_{X_0}}\right)^2}{n_{X_0} - 1}}$$

Sabem per hipòtesis que $\mathbb{P}($ Acceptar $H_1 | H_0$ certa $) = \alpha = 0.05$, és a dir, per rebutjar la hipòtesis nul·la sapiguent que és certa, s'ha de complir que

$$\left| \frac{12 - \hat{X}_0}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_{X_0}^2}{n_{X_0}}}} \right| > |t^k_{1-\frac{\alpha}{2}}|$$

que en ambdues inequacions ens dona

$$n_X > \frac{(t^k_{1-\frac{\alpha}{2}} \theta_0)^2 n_{X_0}}{(12 - \hat{X}_0)^2}$$

Per altra banda, tenim que $\mathbb{P}($ Acceptar $H_0 | H_1$ certa $) = \beta = 0.1$. Per calcular la mida necessària per a que es compleixi aquest succés, suposem que $H_1 : |\theta-\theta_0| = 0.1$, per resoldre un cas senzill. Aleshores, el nostre estadístic quedaria així:

$$\frac{12 - \hat{X}_0 - 0.1}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_{X_0}^2}{n_{X_0}}}}$$

Per tant, per a acceptar la hipòtesis nul·la sapiguent que H_1 és certa, s'ha de complir que

$$\left|\frac{12 - \hat{X}_0 - 0.1}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_{X_0}^2}{n_{X_0}}}}\right| < |t^k_{1-\frac{\alpha}{2}}|$$

Això passa si

$$n_X < \frac{(t^k_{1-\frac{\alpha}{2}})^2(\theta_0^2 + 0.2\theta_0 + 0.01)n_{X_0}}{(11.9 - \hat{X}_0)^2}$$

Aleshores tindrem:

$$\begin{cases} n > \frac{(t^k_{1-\frac{\alpha}{2}}\theta_0)^2 n_{X_0}}{(12-\hat{X}_0)^2} \\ n < \frac{(t^k_{1-\frac{\alpha}{2}})^2(\theta_0^2+0.2\theta_0+0.01)n_{X_0}}{(11.9-\hat{X}_0)^2} \end{cases}$$

El cas $\theta - \theta_0 = -0.1$ és similar.