Estadística

Arnau Mas

2019

Capítol 1

Introducció

S'ha de fer una introducció

Capítol 2

Estimació de paràmetres

Un dels problemes fonamentals de l'estadística és l'obtenció d'informació sobre algun procés tenint accés a un nombre limitat de dades. Situacions d'aquesta mena n'hi ha moltes i molt diverses: enquestes preelectorals, estudis econòmics, experiments científics... Tots aquests casos tenen en comú que l'accés a totes les possibles observacions no és factible, i en segons quins casos impossible. En aquestes circumstàncies apareixen tres preguntes naturals

1. Com podem obtenir resultats útils a partir de les dades disponibles?

2. Quina relació hi ha entre els resultats que obtenim de les dades i el procés en qüestió?

3. Què podem dir amb certesa sobre

La primera pregunta és la que tractem en aquest capítol.

2.1 Mostra i població

La manera de formalitzar la idea d'observacions limitades d'algun fenomen es basa en les eines de la teoria de la probabilitat. A partir d'ara suposarem que tenim una variable aleatòria que segueix una distribució la forma de la qual coneixem, però que depèn d'un o més paràmetres que desconeixem. Escriurem $f_X(x \mid \theta)$[1] per la densitat —o funció de probabilitat si X és discreta— de X, on θ són els paràmetres dels quals pot dependre. Direm que la variable X representa la *població* que estem estudiant.

En general tindrem accés a unes quantes observacions de X, $\{x_1, \cdots x_n\}$. Aquests valors se solen anomenar observacions. Pensarem que aquestes dades o observacions provenen de n variables aleatòries $X_1, \ldots, X_n$ idènticament distribuïdes segons la distribució de X. Si no es diu el contrari, les X_k les suposarem independents —per abreviar farem servir i.i.d.: independents i idènticament distribuïdes—. Diem que el vector $\mathbf{X} = (X_1, \ldots, X_n)$ és una *mostra de mida n*.

[1] La notació fent servir | és comuna en l'estadística Bayesiana. En la comunitat freqüentista és més comuna la notació $f_x(x; \theta)$. La tria de la notació Bayesiana no té, ara per ara, cap significat més enllà de que és la que l'autor considera més estètica.

2.1.1 Mitjana i variància mostrals

Donada una mostra, existeixen diversos cálculs que podem fer per a tenir una idea de com estan distribuïdes les dades. Són les mesures de tendència central i de dispersió tradicionals. Les més bàsiques són la *mitjana mostral*, $\bar{x}$, i la *variància mostral*, s^2:

$$\bar{x} := \frac{1}{n}\sum_{k=1}^{n} x_k,$$
$$s^2 := \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2.$$

De la mateixa manera que pensem en la mostra com un vector aleatori i no només com una seqüència de resultats concrets, també podem introduir la mitjana i variància mostrals com a variables aleatòries, que denotem per $\bar{X}$ i S^2. És a dir

$$\bar{X} := \frac{1}{n}\sum_{k=1}^{n} X_k,$$
$$S^2 := \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2.$$

Tret d'alguns casos concrets com ara la suma de normals independents, calcular la distribució d'una variable aleatòria que és la suma de dues altres variables aleatòries és en general un problema complicat, fins i tot si les variables que sumem són i.i.d.. La distribució de la mitjana serà, doncs, desconeguda en la majoria dels casos. El cas de la variància mostral és encara pitjor ja que ara els sumands ja no són independents. Ara bé teoremes com el Teorema Central del Límit ens permeten conèixer com és la distribució asimptòtica de variables d'aquesta mena, és a dir, quan la mida de la mostra n és gran. El Teorema de Fisher és el que ens dóna la majoria de resultats asimptòtics d'aquest estil. Això té sentit ja que si tenim una mostra petita no esperem poder conèixer amb certesa la distribució de la població, però sí quan la mostra és gran.

El que sí que podem fer, però, és calcular l'esperança i la variància de la mitjana i variància mostrals. Per a la mitjana és un càlcul senzill.

Proposició 2.1. *Per una població X amb $\mathbb{E}[X] = \mu$ i $\mathrm{Var}[X] = \sigma^2$, per una mostra de mida n se satisfà*

(*i*) $\mathbb{E}[\bar{X}] = \mu$,

(*ii*) $\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$.

Demostració. Veure (*i*) és immediat fent servir la linealitat de l'esperança:

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k] = \frac{n\mu}{n} = \mu.$$

Per a (*ii*) podem fer servir que la variància de la suma de variables independents és la suma de les seves variàncies, i que la variància és homogènia de grau 2:

$$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}[X_k] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \qquad \square$$

Aquest resultat tant senzill és la base d'un dels mantres més importants de l'estadística: *els errors van com* $\frac{1}{\sqrt{n}}$.

Capítol 3

Contrast d'hipòtesis

Capítol 4

Intervals de confiança

Capítol 5

Modelització