Capítol 1

INFERÈNCIA, MOSTRATGE I DISTRIBUCIONS MOSTRALS

1.1 Inferència estadística

Per començar anem a definir quin és l'àmbit d'estudi de la inferència estadística des de la seva relació amb el càlcul de probabilitats. El càlcul de probabilitats proporciona una teoria matemàtica que permet analitzar (o modelitzar) les propietats dels fenòmens on intervé l'atzar.
El càlcul de probabilitats utilitza com a model bàsic per a qualsevol situació aleatòria el concepte d'espai de probabilitats $(\Omega, \mathcal{A}, P)$ i una variable aleatòria $X : \Omega \to \mathbb{R}$ definida sobre ell.
El coneixement de la distribució de la variable aleatòria permet:

1. **Anàlisis deductiva de situacions.** *Per exemple: si assumim que el pes dels nadons es distribueix segons una distribució* $N(\mu = 3 \text{ kg}, \sigma = 0.25 \text{ kg})$ *ens pot interessar calcular la probabilitat que un nadó pesi entre* 2.9 *i* 3.1 kg*, o trobar uns valors centrats en la mitjana entre els quals esperem que es trobin el 10% (25%, 50%, 95%,...) dels nadons.*

2. **Modelització de situacions aleatòries.** *Per exemple: si assumim que el temps, en anys, fins que s'espatlla una component d'un ordinador es distribueix segons una distribució exponencial* $T \sim \xi(\lambda = 0.3)$ *ens pot interessar calcular la probabilitat que una component donada duri més de* 4 *anys.*

En els casos anteriors ens trobem en una situació molt usual, on ja disposem d'un model sobre el qual efectuem els càlculs, però del qual desconeixem

la procedència. Sembla raonable, i de fet és precisament així, que si volem adaptar un model a una situació haguem de basar-nos únicament en les observacions del fenomen.
Si volem saber cóm es distribueixen els pesos dels nadons n'agafarem uns quants, els pesarem i després mirarem la distribució d'aquests. Pot ser que no calgui pesar tots els nadons (de fet no és possible!), però tampoc és possible deduir la llei per consideracions purament teòriques.
Ara, enlloc de partir d'un espai de probabilitats partirem d'unes observacions $(x_1, ..., x_n)$ i l'objectiu que perseguirem serà obtenir informació sobre la distribució de probabilitats d'un fenomen a partir d'una observació no exhaustiva d'aquest.

1.2 Problemes d'inferència estadística

Hem presentat com a objectiu de la inferència estadística induir propietats del model probabilístic que representa la població a partir d'un conjunt d'observacions.
Segons el tipus de conclusió que vulguem extreure diferenciarem diferents tipus de problemes:

1. Si volem utilitzar la informació proporcionada per la mostra per obtenir una pronòstic numèric únic (és a dir una única aproximació numèrica) d'una o més característiques de la població tenim un problema d'*estimació puntual*.

2. Si volem obtenir informació sobre un rang de valors dins del qual puguem afirmar, amb un cert grau de confiança, que podem atrapar un paràmetre desconegut de la distribució parlem d'*estimació per interval*.

3. Si el que volem fer és decidir si podem acceptar o hem de rebutjar una afirmació sobre la distribució de probabilitat del fenomen estudiat parlem de *contrast d'hipòtesis*. Aquest contrast pot ser:

    - Paramètric: si l'afirmació (la hipòtesi) es refereix als paràmetres de la distribució.
    - No paramètric: si l'afirmació és sobre la forma de la distribució.

1.3 Distribució de la població

Tot problema d'inferència està motivat per un cert grau de desconeixement de la llei de probabilitats que regeix un determinat fenomen aleatori.

El cas més senzill amb el que ens trobem és quan ens interessa una certa variable X amb una funció de distribució F desconeguda en major o menor grau.
La distribució que teòricament segueix la variable d'interès X en la població rep el nom de *distribució teòrica* o *distribució de la població*. La distribució de la població és important ja que sovint es fa servir per determinar la distribució d'alguna característica dels individus d'una població.
En els models de la inferència estadística indiquem el relatiu grau de desconeixement sobre la distribució F en funció de la seva pertinença a una família $\mathcal{F}$ de distribucions. Per això, enlloc d'explicar que $X \sim F = F_0$ indicarem que $X \sim F \in \mathcal{F}$, on $\mathcal{F}$ pot ser un conjunt més o menys extens de distribucions de probabilitat com totes les *distribucions normals* o les *distribucions simètriques* o les *distribucions discretes sobre* $\mathbb{N}$.
Moltes vegades la distribució poblacional F està completament especificada excepte pel valor d'algun paràmetre o paràmetres. En aquest cas podem concretar més la forma de la família de distribucions:

$$X \sim F \in \mathcal{F} = \left\{ F_\theta : \theta \in \Theta \subset \mathbb{R}^k \right\}$$

on Θ és l'espai dels k paràmetres.
La família de possibles distribucions de probabilitat per a X s'anomena, genèricament, *model estadístic* i s'indica com: $\{X \sim F_\theta : \theta \in \Theta\}$.
Veiem alguns exemples.

Exemple 1.3.1 *Suposem que X representa la durada d'un component electrònic que no envelleix, només s'espatlla. És a dir, si en un instant t està funcionant el seu estat és el mateix que en qualsevol moment del passat i la distribució del temps fins que s'espatlli és la mateixa que al principi. Aquesta propietat s'anomena manca de memòria.*
Un model raonable per aquesta situació el dóna la distribució de Weibull que, en aquest cas, podem definir a través de la següent funció de densitat:

$$f_\theta(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases}$$

La família de distribucions associada és

$$\mathcal{F} = \{F_\theta : \theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty)\}$$

Exemple 1.3.2 *Suposem que volem determinar la massa d'un cert tipus de partícules elementals a partir de les observacions en una cambra de bombolles. En cada observació obtenim una dada de la massa de la partícula*

x_i i associada amb ella un cert error de mesura ε. Si la massa comuna de cadascuna d'elles és μ llavors podem escriure:

$$x_i = \mu + \varepsilon_i \qquad i = 1, ..., n$$

on la distribució $\varepsilon_i \sim F$ és desconeguda. El nostre objectiu és obtenir informació sobre F.
Si admetem que $P(\varepsilon_i < 0) = P(\varepsilon_i > 0)$, segons el grau d'exigència que volem tenir, podem suposar:

- *Amb un enfocament d'*inferència paramètrica*:*

$$X \sim F \in \mathcal{F} = \{N(0, \sigma) : \sigma \in \mathbb{R}^+\}$$

- *Amb un enfocament d'*inferència no paramètrica*:*

$$X \sim F \in \mathcal{F} = \{\text{Distribucions simètriques}\}$$

1.4 Mostra aleatòria simple

1.4.1 Definició

Per estudiar un problema d'inferència estadística analitzem una mostra de mida n. Es tracta d'escollir n individus o elements de la població Ω

$$\omega_1, \omega_2, \ldots, \omega_n$$

que siguin representatius. El valor de n i la forma d'elecció dels individus de la mostra és una materia d'Estadística anomenada *Mostratge estadístic*. Per ara i per simplificar, només cal dir que l'elecció es fa de forma que tots els individus tenen la mateixa probabilitat de ser presents a la mostra, si cal amb reemplaçament, i que el valor de n està donat.
En realitat, el que ens interessa veritablement, no són els individus de la mostra sinó les mesures d'una característica X sobre ells. És a dir, els valors d'una variable aleatòria X sobre aquests individus

$$X(\omega_1) = x_1, X(\omega_2) = x_2, \ldots, X(\omega_n) = x_n$$

També podem pensar que els valors mostrals $x_1, x_2, \ldots, x_n$ són generats directament des de la variable aleatòria. En tot cas, els valors mostrals no són únics i podem generar vàries mostres

$$\begin{array}{ccccc} x_1^1 & x_2^1 & x_3^1 & \ldots & x_n^1 \\ x_1^2 & x_2^2 & x_3^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & & \vdots \\ x_1^s & x_2^s & x_3^s & \ldots & x_n^s \end{array}$$

Si tots els valors són independents, de la mateixa forma que $x_1, x_2, x_3, \ldots, x_n$ és una mostra generada per X, podem considerar tots els $x_1^i \quad i = 1, ..., s$ provinents d'una variable aleatòria X_1 amb la mateixa distribució que X $X_1 \stackrel{d}{=} X$ i que genera els primers valors, els x_i^2 provinents d'una variable aleatòria $X_2 \stackrel{d}{=} X$ que genera els segons i així successivament.
Tot això ens porta a definir el concepte de mostra aleatòria d'una forma molt convenient per treballar amb ella:

Definició 1.1 *Una mostra aleatòria simple de mida n d'una variable aleatòria X amb distribució F és una collecció de n variables aleatòries independents $X_1, X_2, \ldots, X_n$ amb la mateixa distribució F que X. Això se sol indicar com:*

$$\mathbf{X} = X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} X$$

Definició 1.2 *El conjunt $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ d'observacions concretes de $X_1, X_2, \ldots, X_n$ s'anomena* realització *de la mostra.*

1.4.2 Distribució de la mostra

Una mostra aleatòria simple, com a vector aleatori n-dimensional que és, té una distribució conjunta o *distribució de la mostra* que depèn de F, però que òbviament és diferent, ja que en particular X i $\mathbf{X}$ tenen dimensions diferents. Ara bé, gràcies a la independència de les variables $X_1, X_2, \ldots, X_n$, la funció de distribució conjunta de $\mathbf{X}$, que podria ser molt complicada, pren una forma molt senzilla. En resum:

Definició 1.3 *S'anomena distribució de la mostra d'una variable aleatòria $X \sim F$ a la distribució del vector aleatori n-dimensional $(X_1, X_2, \ldots, X_n)$*

$$G(x_1, x_2, ..., x_n) = F(x_1)F(x_2)\cdots F(x_n)$$

En els casos particulars en que X sigui discreta o absolutament contínua la distribució conjunta de la mostra sol expressar-se mitjançant la funció de massa de probabilitat o la funció de densitat:

- Per a variables discretes:

$$\begin{aligned} p_G(x_1, x_2, \ldots, x_n) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P\left(X = x_i\right) = \prod_{i=1}^{n} p_F\left(x_i\right), \end{aligned}$$

- Per a variables absolutament contínues:

$$g(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

Exemple 1.4.1 *Una moneda té una probabilitat θ de sortir cara. Volem estudiar la variable aleatòria:*

$$X = \begin{cases} 1 & \textit{si surt cara} \\ 0 & \textit{si surt creu} \end{cases}$$

amb densitat $P\{X = 1\} = \theta$, $P\{X = 0\} = 1 - \theta$. És a dir

$$X \sim F_\theta \in \mathcal{F} = \{F_\theta = B(1, \theta) : \theta \in (0, 1)\}$$

Suposem que fem tres llançaments. Les possibles mostres són:

X_1	X_2	X_3	*Probabilitat*
1	1	1	θ^3
1	0	0	$\theta(1-\theta)^2$
0	1	0	$\theta(1-\theta)^2$
0	0	1	$\theta(1-\theta)^2$
1	0	1	$\theta^2(1-\theta)$
1	1	0	$\theta^2(1-\theta)$
0	1	1	$\theta^2(1-\theta)$
0	0	0	$(1-\theta)^3$

El mostratge ha especificat la distribució conjunta de la mostra a través de la distribució desconeguda F_θ. Si escrivim la funció de probabilitats de la variable aleatòria com $f_\theta(x) = \theta^x(1-\theta)^{1-x}$, llavors la funció de probabilitats de la mostra la podem expressar com:

$$g_\theta(x_1, x_2, x_3) = \theta^{x_1+x_2+x_3}(1-\theta)^{3-(x_1+x_2+x_3)}$$

1.5 Estadístics

1.5.1 Definició

Per aconseguir l'objectiu de realitzar inferències sobre la població a partir de la mostra ens solem basar en la realització de càlculs sobre la mostra per mirar d'obtenir la informació que desitgem. En aquest procés apareixen els conceptes d'estadístic i el cas particular, que més ens interessa a nosaltres, d'estimador. Un estadístic és una funció de la mostra que no depèn del valor del paràmetre.

Definició 1.4 *Donada una mostra aleatòria simple $X_1, X_2, \ldots, X_n$ i una funció mesurable $T : \mathbb{R}^n \longrightarrow \mathbb{R}^k$, llavors $T(X_1, X_2, \ldots, X_n)$ és un vector aleatori (variable aleatòria quan $k = 1$). Si T no depèn de θ (on θ és un paràmetre a especificar en F_θ), llavors T rep el nom d'estadístic.*

Només pel seu nom, sembla evident que un estimador d'un paràmetre θ serà alguna funció de la mostra que serveixi per aproximar, en algun sentit, el valor desconegut de θ. Si afegim la condició raonable que un estimador no pugui prendre valors que no pot prendre el paràmetre podem donar la següent definició.

Definició 1.5 *Un estimador d'un paràmetre θ és un estadístic T el recorregut del qual és l'espai dels paràmetres, és a dir:*

$$\begin{array}{rccc} T: & \mathbb{R}^n & \longrightarrow & \mathbb{R}^k \\ & (x_1, x_2, \ldots, x_n) & \longrightarrow & (t_1, \ldots, t_k) \quad \in \Theta \subset \mathbb{R}^k \end{array}$$

1.5.2 Distribució en el mostratge d'un estadístic

Donat un estadístic $T(X_1, X_2, \ldots, X_n)$ ens interessa conèixer la seva distribució de probabilitat, ja que per fer inferència ens caldrà fer càlculs del tipus

$$P\left[T(X_1, X_2, \ldots, X_n) > t_0\right]$$

La distribució de probabilitat de l'estadístic s'anomena distribució mostral o distribució en el mostratge de l'estadístic. Trobar-la és un problema que pot ser des de bastant senzill fins a extremadament complicat. Algunes de les tècniques utilitzades per mirar de resoldre'l són les següents:

- Ús de la tècnica de canvi de variable.
- Ús de la funció generatriu de moments.
- Aplicació del Teorema Central del Límit.

Exemple 1.5.1 *Sigui $X \sim F_\theta$ una variable aleatòria absolutament contínua amb densitat*

$$f_\theta(x) = e^{-(x-\theta)} e^{-e^{-(x-\theta)}} \quad \theta \in \mathbb{R}$$

i considerem l'estadístic

$$T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} e^{-X_i}$$

Si apliquem el teorema de canvi de variable unidimensional, s'obté fàcilment que la variable aleatòria $Y = e^{-X}$ segueix una distribució exponencial de paràmetre $e^{-\theta}$, d'on la suma seguirà una distribució gamma $T \sim \Gamma(e^{-\theta}, n)$.

Exemple 1.5.2 *Suposem que X representa el nombre d'avaries en una màquina al cap d'un mes. Aquest valor varia mes a mes. Sigui $\bar{X}$ la mitjana d'avaries en n mesos. Si X segueix una distribució de Poisson $P(\lambda)$, quina és la distribució de $\bar{X}$?*
Com que la suma de Poisson i.i.d. és $\sum_{i=1}^{n} X_i \sim P(n\lambda)$

$$P\left[\bar{X} = r\right] = P\left[\sum_{i=1}^{n} X_i = nr\right] = \frac{e^{-n\lambda}\left(n\lambda\right)^{nr}}{(nr)!}$$

Com passa en aquest exemple, un dels estadístics pel qual sovint desitgem calcular la distribució en el mostratge és la mitjana aritmètica. Una manera útil de fer-ho és amb la funció generatriu de moments i l'aplicació del següent lema.

Lema 1 *Si X és una v.a. amb $M_X(t)$ com funció generatriu de moments, llavors la f.g.m. de $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ és*

$$M_{\bar{X}_n}(t) = \left[M_X(t/n)\right]^n$$

Demostració:
La demostració és immediata a partir de la definició o per les propietats de la funció generatriu de moments.
Si apliquem directament la definició de la f.g.m tenim:

$$\begin{aligned} E(e^{t\bar{X}_n}) = E(e^{t\frac{1}{n}\sum_{i=1}^{n} X_i}) = E\left(\prod_{i=1}^{n} e^{\frac{t}{n}X_i}\right) = \prod_{i=1}^{n} E\left(e^{\frac{t}{n}X_i}\right) \\ = \prod_{i=1}^{n} M_{X_i}(t/n) = \left[M_X(t/n)\right]^n \end{aligned}$$

Si fem servir les propietats de la f.g.m. tenim:

1. Donat que $M_{aX}(t) = M_X(at)$ i si $a = \frac{1}{n}$, llavors $M_{\bar{X}}(t) = M_{\sum_{i=1}^{n} X_i}(t/n)$.
2. $M_{\sum_{i=1}^{n} X_i}(t/n) \overset{\text{ind}}{=} \prod_{i=1}^{n} M_{X_i}(t/n) \overset{\text{id}}{=} \left[M_X(t/n)\right]^n$.

Exemple 1.5.3 *Per a una variable aleatòria $X \sim N(\mu, \sigma)$ i per tant $M_X(t) = \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right)$, aleshores*

$$\begin{aligned} M_{\bar{X}_n}(t) &= \left[\exp\left(\frac{t\mu}{n} + \frac{t^2\sigma^2}{n^2 2}\right)\right]^n \\ &= \exp\left[n\left(\frac{t\mu}{n} + \frac{t^2\sigma^2}{n^2 2}\right)\right] \\ &= \exp\left[t\mu + \frac{1}{2}t^2\left(\frac{\sigma}{\sqrt{n}}\right)^2\right] \end{aligned}$$

que és la funció generatriu de moments d'una variable $N(\mu, \sigma/\sqrt{n})$.

1.6 La distribució empírica

1.6.1 Definició

En l'apartat anterior hem vist que a partir d'una mostra $X_1, X_2, \ldots, X_n$ té interès considerar la distribució mostral com la distribució conjunta del vector aleatori $(X_1, X_2, \ldots, X_n)$, sense que intervingui una realització concreta de la mostra $x_1, x_2, \ldots, x_n$. Un enfocament diferent consisteix en associar una distribució particular directament a les observacions $x_1, x_2, \ldots, x_n$ amb la pretensió que, en tant que la mostra "representa" la v.a. X, aquesta distribució associada a la mostra $F_n(x)$ emuli la distribució de la població. Aquesta distribució s'anomena distribució empírica o distribució mostral i es defineix així:

$$F_n(x) = \frac{k(x)}{n}$$

on $k(x)$ és el nombre de dades mostrals menors o iguals que x. A la pràctica es construeix per ordenació de la mostra

$$x_1, x_2, \ldots, x_n \longrightarrow x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

i amb la següent definició:

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \frac{k}{n} & \text{si } x_{(k)} \leq x < x_{(k+1)} \\ 1 & \text{si } x_{(n)} \leq x \end{cases}$$

Exemple 1.6.1 *Extraïem una mostra i obtenim:*

x_1	x_2	x_3	x_4	x_5	x_6	x_7
5.1	3.4	1.2	17.6	2.1	16.4	4.3

Un cop ordenada ens queda:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$
x_3	x_5	x_2	x_7	x_1	x_6	x_4
1.2	2.1	3.4	4.3	5.1	16.4	17.6

i si fem la representació gràfica:

La distribució empírica reflecteix exclusivament els valors observats a la mostra i per tant no es relaciona directament ni amb la distribució conjunta de la mostra $G(x_1, x_2, \ldots, x_n)$ ni amb la distribució de la població F. Malgrat això, com és raonable esperar, $F_n(x)$ proporciona una imatge aproximada de la distribució de la població d'on s'ha extret la mostra.

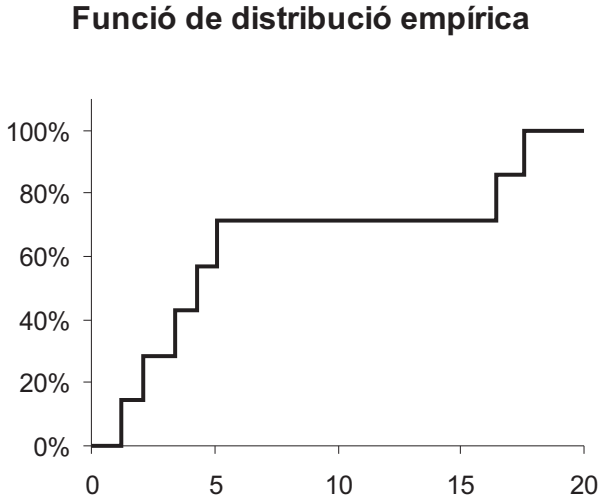

Figura 1.1: Funció de distribució empírica amb les dades de l'exemple

1.6.2 Relació entre la distribució empírica i la poblacional

L'estudi de la relació entre $F_n(x)$ i $F(x)$ es pot fer des de diversos punts de vista. Podem considerar $F_n(x)$ com un estadístic o com una distribució.
Si considerem que $F_n(x)$ representa la freqüència relativa en la mostra de l'esdeveniment $[X \leq x]$ que té probabilitat $F(x)$, llavors és un estadístic. Aleshores, té sentit que considerem la seva distribució en el mostratge, $P[F_n(x) \leq z]$ i que estudiem els moments d'aquesta distribució. En aquest cas també té sentit aplicar les lleis dels grans nombres i determinar sota quines condicions es verifica que $F_n(x) \longrightarrow F(x)$ en probabilitat o quasi-segurament.
Si considerem que $F_n(x)$ representa directament una distribució de probabilitat, definida sobre el conjunt $\{(x_1, x_2, \ldots, x_n)\}$ té sentit que estudiem els seus moments, és a dir, els de la variable que la té per distribució. Si la tractem com una distribució de probabilitat també té sentit estudiar la seva convergència en distribució.

L'estadístic $F_n(x)$

Per operar més fàcilment amb ella, podem escriure $F_n(x)$ com:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{[X_i \leq x]}(x) = \frac{1}{n}\sum_{i=1}^{n}W_i(x)$$

i per simplificar la notació posarem W_i enlloc de $W_i(x)$.
En l'exemple anterior, si $x = 10$ tenim $W_1 = W_2 = W_3 = W_5 = W_7 = 1$, $W_4 = W_6 = 0$ i per tant $F_n(10) = 5/7$.
Tal com hem definit W_i, aquest val 1 si $X_i \leq x$ i 0 si $X_i > x$, és a dir

$$W_i = \begin{cases} 1 & \text{amb } P[W_i = 1] = P[X_i \leq x] = F(x) \\ 0 & \text{amb } P[W_i = 0] = 1 - P[X_i \leq x] = 1 - F(x) \end{cases}$$

de forma que resulta clar que

$$W_i \sim B(1, F(x)) \quad \text{de manera que} \quad \sum_{i=1}^{n}W_i \sim B(n, F(x))$$

i la variable $\frac{1}{n}\sum_{i=1}^{n}W_i$ pren els valors $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$ amb probabilitats definides per la distribució binomial $B(n, F(x))$.
De la representació anterior és immediat que si posem $Y \sim B(n, p)$ on $p = F(x)$, tenim:

$$E(F_n(x)) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{1}{n}np = p = F(x)$$
$$\text{var}(F_n(x)) = \text{var}\left(\frac{Y}{n}\right) = \frac{1}{n^2}\text{var}(Y) = \frac{npq}{n^2} = \frac{pq}{n} = \frac{F_n(x)\cdot(1 - F_n(x))}{n^2}$$

La v.a. $F_n(x)$ és un estadístic que pren valors sobre el mateix conjunt que $F(x)$ i, per tant, és un estimador en el sentit definit més amunt. És d'esperar que $F_n(x)$ s'apropi a $F(x)$ en algun sentit. Per aplicació de les lleis dels grans nombres, i un cop vista la representació de $F_n(x)$ com un promig de variables i.i.d. $B(1, p)$, és immediat que $F_n(x) \longrightarrow F(x)$ en probabilitat i quasi-segurament.
El resultat anterior es reforça si estudiem l'aproximació de $F_n(x)$ a $F(x)$ a través de l'estadístic de Kolmogorov

$$D_n = \sup_x |F_n(x) - F(x)|$$

Es demostra[1] (Teorema de Glivenko-Cantelli) que $F_n(x)$ convergeix a $F(x)$ quasi segurament i uniformement en x, és a dir

$$P\left[\lim D_n = 0\right] = 1$$

Pel que fa a la convergència en distribució de $F_n(x)$ podem enunciar la següent propietat.

Proposició 1 *Sigui $x_1, x_2, \ldots, x_n$ una realització d'una mostra aleatòria simple de la distribució F i sigui $F_n(x)$ la seva funció de distribució empírica. L'estadístic $F_n(x)$ té una distribució de probabilitat asimptòticament normal*

$$F_n(x) \sim AN\left(F(x), \sqrt{\frac{F(x)\cdot(1-F(x))}{n}}\right)$$

La demostració és immediata si considerem la representació de $F_n(x)$ com suma de variables aleatòries i.i.d. i apliquem el Teorema Central del Límit.

La distribució de probabilitat $F_n(x)$

Si considerem $F_n(x)$ com una distribució de probabilitat, té sentit estudiar els moments de la variable que té a $F_n(x)$ per funció de distribució, així com la seva convergència en distribució. Com a conseqüència d'alguns dels resultats anteriors, $F_n(x)$ convergeix en distribució cap a la distribució poblacional. Es pot justificar simplement tenint en compte que la convergència en probabilitat implica la convergència en distribució.
Pel que fa als moments de la distribució empírica la següent secció estudia els moments mostrals que són, de fet, els moments d'una variable aleatòria que tingui a $F_n(x)$ com funció de distribució.

1.7 Els moments mostrals

1.7.1 Definició

Sigui F_n la v.a. que té $F_n(x)$ per distribució. La funció de densitat de probabilitat de F_n és una densitat discreta que assigna probabilitats $1/n$ a cadascuna de les observacions mostrals $x_1, x_2, \ldots, x_n$. Així doncs, té sentit

[1] Veieu: *Estadística.* Fortiana, J. i Nualart, D. Publicacions U.B.

que calculem els seus moments que es coneixen com a *moments mostrals* a_k i també els seus *moments mostrals centrats respecte la mitjana* b_k.

$$a_k = E(F_n^k) = \sum_{i=1}^{n} x_i^k \cdot P(F_n = x_i) = \sum_{i=1}^{n} x_i^k \cdot \frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Observem que dues mesures conegudes de l'estadística descriptiva adquireixen un significat diferent:

- Mitjana mostral = Mitjana de la distribució mostral

$$a_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variància mostral = Variància de la distribució mostral

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

1.7.2 Distribució en el mostratge dels moments mostrals

Donada una m.a.s. $X_1, X_2, \ldots, X_n$, els moments mostrals són estadístics i, com a tals, tenen la seva distribució en el mostratge. Per exemple, $a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.
La distribució en cada cas pot ser complexa i dependre de la distribució poblacional subjacent.
El que sí és possible calcular són els moments dels moments mostrals o, més ben dit, els moments de les distribucions en el mostratge dels moments mostrals.

1. Si considerem $a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ i escrivim $\alpha_k = E(X^k)$ com el moment poblacional d'ordre k, tenim:

$$E(a_k) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n}\cdot n \cdot \alpha_k = \alpha_k$$

$$\begin{aligned}\text{var}(a_k) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}\left(X_i^k\right) = \frac{1}{n}\text{var}(X^k) \\ &= \frac{1}{n}\left[E(X^{2k}) - (E(X^k))^2\right] = \frac{\alpha_{2k} - \alpha_k^2}{n}\end{aligned}$$

2. Si considerem $s^2 = b_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$, podem calcular:

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\bar{X})^2 = \alpha_2 - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\sigma^2$$

El càlcul de la variància de s^2 és feixuc[2] i no el farem aquí. El seu valor és

$$\text{var}(s^2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\,\mu_2^2)}{n^2} + \frac{\mu_4 - 3\,\mu_2^2}{n^3}$$

on μ_k és el moment poblacional centrat d'ordre k.

1.7.3 Propietats asimptòtiques dels moments mostrals

Convergència en probabilitat

Els moments mostrals, tan respecte l'origen com respecte la mitjana, convergeixen cap als moments poblacionals. És possible establir la convergència en base a la llei forta dels grans nombres (convergència quasi-segura) o a la llei feble (convergència en probabilitat). Si ens limitem a aquesta darrera podem afirmar que

$$a_k \xrightarrow{P} \alpha_k \quad \text{és a dir} \quad \lim_{n\to\infty} P[|a_k - \alpha_k| \geq \epsilon] = 0$$

La prova es basa en la desigualtat de Txebitxev. Si suposem que $\alpha_{2k} < \infty$, tenim

$$P[|a_k - \alpha_k| \geq \epsilon] \leq \frac{E|a_k - \alpha_k|^2}{\epsilon^2} = \frac{\text{var}(a_k)}{\epsilon^2} = \frac{\alpha_{2k} - \alpha_k^2}{n\epsilon^2} \longrightarrow 0$$

Aquesta propietat és important perquè farà possible el concepte d'estimador consistent i en ella es fonamenta un mètode d'estimació anomenat mètode dels moments.

Distribució asimptòtica

Si considerem el moment mostral $a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, llavors $n \cdot a_k$ és una suma de variables aleatòries i.i.d. a la que li podem aplicar el Teorema Central del Límit. Tal i com hem vist:

$$E(na_k) = n\alpha_k \qquad \text{var}(na_k) = n^2\text{var}(a_k) = n^2\,\frac{\alpha_{2k} - \alpha_k^2}{n}$$

[2]Veieu: *Métodos matemáticos de la estadística*, d'H. Cramer. Ed. Aguilar

i pel Teorema Central del Límit de Lindeberg-Levy la variable

$$\frac{na_k - E(na_k)}{\sqrt{\text{var}(na_k)}} = \frac{na_k - n\alpha_k}{n\sqrt{\text{var}(a_k)}} = \frac{a_k - \alpha_k}{\sqrt{\text{var}(a_k)}}$$

verifica

$$\frac{a_k - \alpha_k}{\sqrt{\text{var}(a_k)}} \xrightarrow{\mathcal{L}} N(0,1)$$

és a dir

$$a_k \sim AN\left(\alpha_k, \sqrt{\frac{\alpha_{2k} - \alpha_k^2}{n}}\right)$$

1.8 Mostratge en poblacions normals

Com hem vist, a partir d'una m.a.s. $X_1, X_2, \ldots, X_n$ i si considerem un estadístic $T(X_1, X_2, \ldots, X_n)$, pot resultar complicat obtenir la seva distribució en el mostratge. Aquesta distribució depèn de:

- La forma funcional de $T(X_1, X_2, \ldots, X_n)$.
- La distribució subjacent de X, és a dir, la distribució de la població.

Hi ha un cas especial en que el problema s'ha estudiat en profunditat per a alguns estadístics de gran importància pràctica. Si $X \sim N(\mu, \sigma)$ és possible trobar la distribució dels estadístics més utilitzats com $\bar{X}$ i $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$. De fet obtindrem la distribució de funcions d'aquests estadístics com

$$\frac{\bar{X} - \mu}{s/\sqrt{n-1}}; \quad \frac{ns^2}{\sigma^2}; \quad \bar{X}_1 - \bar{X}_2; \quad \frac{S_1^2/(n_1-1)}{S_2^2/(n_2-1)}$$

on $s^2 = (1/n)S^2$.
En l'estudi de les distribucions d'aquests estadístics apareixen algunes distribucions de probabilitat que han resultat ser de gran utilitat. Són les anomenades "distribucions derivades de la normal" i es coneixen pel nom de l'investigador que les va formular:

- la χ^2 khi-quadrat de Pearson
- la t de Student (Gosset)
- la F de Fisher-Snedecor

1.8.1 La distribució khi-quadrat

Siguin $X_1, X_2, \ldots, X_k$ un conjunt de v.a. independents sobre un mateix espai de probabilitat $(\Omega, \mathcal{A}, P)$ i amb distribució comuna $N(0,1)$. Considerem la variable

$$Y = X_1^2 + X_2^2 + \cdots + X_k^2$$

La distribució de la variable Y s'anomena khi-quadrat amb k graus de llibertat.
La funció de densitat de la variable aleatòria Y és

$$f(x) = \frac{1}{\Gamma(k/2)2^{k/2}} e^{-x/2} x^{k/2-1} \qquad \text{si } x > 0$$

De manera que resulta que $Y = \sum_{i=1}^{k} X_i^2$ té una distribució gamma $G\left(\frac{1}{2}, \frac{k}{2}\right)$ i la seva f.g.m. és

$$M(t) = (1-2t)^{-k/2} \quad \text{si } t < 1/2$$

Propietats

1. Si recordem que per a $X \sim G(p, \alpha)$ aleshores $E(X) = \frac{p}{\alpha}$ i $\text{var}(X) = \frac{p}{\alpha^2}$, resulta
$$E(Y) = \frac{k/2}{1/2} = k \qquad \text{var}(Y) = \frac{k/2}{1/4} = 2k$$

2. De l'additivitat (reproductivitat) de les lleis gamma es dedueix també la reproductivitat de la khi-quadrat χ^2, és a dir
$$Y_1^2 \sim \chi_{n_1}^2,\ Y_2^2 \sim \chi_{n_2}^2 \quad \text{indep.} \longrightarrow Y_1^2 + Y_2^2 \sim \chi_{n_1+n_2}^2$$

3. Com Y és la suma de v.a. independents $X_i^2 \sim \chi_1^2$ es verifica
$$\frac{Y-k}{\sqrt{2k}} \xrightarrow{\mathcal{L}} N(0,1)$$
Però és millor l'aproximació de Fisher
$$\sqrt{2\chi_k^2} - \sqrt{2k-1} \xrightarrow{\mathcal{L}} N(0,1)$$
d'on s'obté per valors de $k \geq 30$
$$\chi_k^2 \overset{\text{aprox}}{=} \frac{1}{2}(Z + \sqrt{2k-1})^2$$
on $Z \sim N(0,1)$.

1.8.2 Distribució t de Student

Siguin Y, Z dues variables aleatòries independents amb distribucions $Z \sim N(0,1)$ i $Y \sim \chi^2_m$, llavors es diu que la variable aleatòria

$$t = \frac{Z}{\sqrt{Y/m}}$$

té una distribució t de Student amb m graus de llibertat.
La seva funció de densitat és

$$f(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}}\left(1+\frac{t^2}{m}\right)^{-(m+1)/2} \qquad t \in \mathbb{R}$$

Aquesta expressió s'obté de la resolució del corresponent problema de canvi de variable per trobar la distribució d'un quocient.
Es tracta d'una distribució unimodal i simètrica respecte el zero. La distribució depèn de m, que anomenem els *graus de llibertat* (g.ll.). A mida que m creix, la forma acampanada es va "tancant", acostant-se a la llei normal:

$$\left(1+\frac{t^2}{m}\right)^{-(m+1)/2} \xrightarrow{m\to\infty} e^{-t^2/2}$$

Aquest fet és molt rellevant en inferència estadística.

Propietats

1. Si $m = 1$, aleshores la t és una Cauchy i, en particular, no té esperança.
2. Per a $m > 1$, $E(t) = 0$ i per $m > 2$, $\text{var}(t) = m/(m-2)$.
3. Quan $m \to \infty$, llavors $t \xrightarrow{P} N(0,1)$.

1.8.3 La distribució F de Fisher

Aquesta distribució apareix quan es considera un quocient entre dues distribucions khi-quadrat $U \sim \chi^2_m$,$V \sim \chi^2_n$ amb m i n g.ll. respectivament. En concret diem que la variable aleatòria

$$F = \frac{U/m}{V/n}$$

segueix una distribució F de Fisher amb m i n graus de llibertat. La funció de densitat té la forma:

$$f(x) = \frac{m^{m/2}n^{n/2}\Gamma\left[(m+n)/2\right]}{\Gamma(m/2)\,\Gamma(n/2)} \cdot \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}} \qquad \text{per a } x > 0$$

Propietats

1. L'esperança i la variància són

$$E(F) = \frac{n}{n-2} \qquad \text{var}(F) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

2. Aquesta distribució té una moda en $x = \frac{m-2}{m} \cdot \frac{n}{n+2}$, sempre que $m > 2$.

3. Si $F \sim F_{m,n}$, aleshores resulta que $1/F \sim F_{n,m}$ i llavors:

$$P(F \leq x) = P\left(\frac{1}{F} \geq \frac{1}{x}\right) = 1 - P\left(\frac{1}{F} \leq \frac{1}{x}\right)$$

   Aquesta propietat és de gran utilitat en l'ús de les taules.

4. Quan $n \to \infty$, $F_{m,\infty} \xrightarrow{\mathcal{L}} \chi^2_m$.

5. Quan $m \to \infty$ i $n \to \infty$, $F_{m,n} \xrightarrow{P} 1$.

1.8.4 Distribució de la mitjana i la variància mostrals

Si la distribució de la població d'on prové la mostra és una llei normal és possible calcular exactament la distribució en el mostratge d'alguns estadístics que són molt importants en inferència estadística. Els principals resultats d'aquesta secció van ser obtinguts per Fisher. La presentació que farem aquí, treta de Rohatgi (1975), evita haver de fer servir resultats algebraics i sobre distribució de formes quadràtiques.
El teorema que enunciem i demostrem a continuació diu simplement que donada una mostra aleatòria simple d'una distribució $N(\mu, \sigma)$, els estadístics $\bar{X}$ i $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ són independents. Com s^2 és funció de $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ se'n deduirà de forma immediata la independència entre $\bar{X}$ i s^2 i de forma senzilla la distribució d'algunes funcions d'estadístics molt utilitzats en inferència estadística.
La demostració del teorema consisteix en calcular la funció generatriu de moments conjunta de

$$(\bar{X}, X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$$

i veure que coincideix amb el producte de les funcions generatrius de

$$\bar{X} \text{ i de } (X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$$

Teorema 1.1 *Sigui $X_1, X_2, \ldots, X_n$ una m.a.s. d'una variable $X \sim N(\mu, \sigma)$, aleshores $\bar{X}$ i $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ són independents.*

Demostració:
Sabem que la funció generatriu de moments d'una variable aleatòria normal $N(\mu, \sigma)$ és

$$M_X(t) = E \exp(Xt) = \exp\{\mu t + (\sigma t)^2/2\}$$

Anem a calcular la funció generatriu de moments conjunta de $\bar{X}$ i $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$:

$$\begin{aligned}
M(t, t_1, t_2, \ldots, t_n) &= E \exp\left\{t\bar{X} \right. \\
&\qquad \left. + t_1(X_1 - \bar{X}) + t_2(X_2 - \bar{X}) + \cdots + t_n(X_n - \bar{X})\right\} \\
&= E \exp\left\{\sum_{i=1}^{n} t_i X_i - \left(\sum_{i=1}^{n} t_i - t\right)\bar{X}\right\} \\
&= E \exp\left\{\sum_{i=1}^{n} X_i \left(t_i - \frac{\sum_{i=1}^{n} t_i - t}{n}\right)\right\} \\
&= E\left[\prod_{i=1}^{n} \exp\left\{\frac{X_i\left(nt_i - n\bar{t} + t\right)}{n}\right\}\right] \\
&= \prod_{i=1}^{n} E \exp\left\{\frac{X_i\left[t + n\left(t_i - \bar{t}\right)\right]}{n}\right\} \\
&= \prod_{i=1}^{n} \exp\left\{\frac{\mu\left[t + n\left(t_i - \bar{t}\right)\right]}{n} + \frac{\sigma^2}{2}\frac{1}{n^2}\left[t + n\left(t_i - \bar{t}\right)\right]^2\right\} \\
&= \exp\left\{\frac{\mu}{n}\left[nt + n\sum_{i=1}^{n}\left(t_i - \bar{t}\right)\right] + \frac{\sigma^2}{2n^2}\sum_{i=1}^{n}\left[t + n\left(t_i - \bar{t}\right)\right]^2\right\} \\
&= \exp\left(\mu t\right) \exp\left\{\frac{\sigma^2}{2n^2}\left(nt^2 + n^2\sum_{i=1}^{n}\left(t_i - \bar{t}\right)^2\right)\right\} \\
&= \exp\left(\mu t + \frac{\sigma^2}{2n}t^2\right) \exp\left\{\frac{\sigma^2}{2}\left(\sum_{i=1}^{n}\left(t_i - \bar{t}\right)^2\right)\right\} \\
&= M_{\bar{X}}(t) \cdot M_{(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})}(t_1, t_2, \ldots, t_n) \\
&= M(t, 0, \ldots, 0) \cdot M(0, t_1, t_2, \ldots, t_n)
\end{aligned}$$

En aquest càlcul fem servir $\bar{t} = (1/n)\sum_{i=1}^{n} t_i$ i el fet que $\sum_{i=1}^{n}(t_i - \bar{t}) = 0$.

Els següents corol.laris del teorema anterior posen de manifest la seva aplicabilitat.

Siguin

$$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad s^2 = \frac{1}{n} S^2 \qquad \hat{s}^2 = \frac{1}{n-1} S^2$$

Corollari 1 *$\bar{X}$ i S^2 són independents. $\bar{X}$ i s^2 són independents. $\bar{X}$ i $\hat{s}^2$ són independents.*

Corollari 2 *Amb σ^2 coneguda, l'estadístic*

$$\frac{S^2}{\sigma^2} = \frac{ns^2}{\sigma^2} = \frac{(n-1)\hat{s}^2}{\sigma^2}$$

es distribueix com una χ^2_{n-1} amb $n-1$ graus de llibertat.

Corollari 3 *Amb μ coneguda, la distribució de l'estadístic*

$$\frac{\bar{X} - \mu}{s}\sqrt{n-1} = \frac{\bar{X} - \mu}{\hat{s}}\sqrt{n}$$

és una t de Student amb $n-1$ graus de llibertat.

Corollari 4 *Donades dues mostres aleatòries simples de dues poblacions normals $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ respectivament i preses de forma independent, l'estadístic (amb les variancies conegudes)*

$$\frac{\hat{s}_1^2/\sigma_1^2}{\hat{s}_2^2/\sigma_2^2}$$

es distribueix com una F de Fisher amb $n_1 - 1, n_2 - 1$ graus de llibertat. En el cas particular de que les variancies de les poblacions siguin iguals, és a dir $\sigma_1^2 = \sigma_2^2$, llavors

$$\frac{\hat{s}_1^2}{\hat{s}_2^2}$$

es distribueix com una F de Fisher amb $n_1 - 1, n_2 - 1$ graus de llibertat.

Corollari 5 *Donades dues mostres aleatòries simples*

$$X_1, X_2, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1) \qquad Y_1, Y_2, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2)$$

preses de forma independent, l'estadístic (amb els paràmetres coneguts)

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{(n_1 - 1)\hat{s}_1^2/\sigma_1^2 + (n_2 - 1)\hat{s}_2^2/\sigma_2^2} \sqrt{\frac{n_1 + n_2 - 2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

segueix una distribució t de Student amb $n_1 + n_2 - 2$ graus de llibertat.
En el cas particular de que $\sigma_1^2 = \sigma_2^2$, les variancies se simplifiquen i desapareixen de l'estadístic anterior.