APUNTE ESTADISTICA
Departamento de Ingeniería Matemática

INFERENCIA ESTADISTICA

N.Lacourly, L.Cerda, L.Bruna y R.Abt

1 INTRODUCCION

Problema: Un agricultor puede plantar papas o tomates. Si planta papas y llueve, el agricultor gana solo $600.000, pero si no llueve gana $1.800.000. En cambio, si planta tomates y llueve, gana $1.200.000, pero si no llueve, gana solo $800.000. Qué le conviene plantar al agricultor?. Para analizar el problema, debemos identificar los elementos del mismo. En primer lugar tenemos:

- Dos acciones: sembrar papas o sembrar tomates.
- El clima: llueve o no llueve.
- Los pagos asociados a las combinaciones de ambos elementos

Lo cual se pueder resumir en el siguiente cuadro:

	P	T
LL	600	1.200
N-LL	1.800	800

Siendo P: plantar papas, T: plantar tomates, LL: llueve y N-LL: no llueve. (Pagos en miles de pesos).

COMO DECIDIR?

- Para que el agricultor decida, debe contar con un criterio o regla, que basándose en la información disponible, le permita obtener el mejor resultado. En este caso, el agricultor busca ganar el mayor beneficio con la venta.
- Cuando NO existe incertidumbre, la selección de un criterio puede obedecer a la actitud frente al riesgo que presenten los tomadores de decisión. Algunos de estos criterios son: Maximax: el máximo de los máximos(para optimistas), Maximin: el máximo de los mínimos (para pesimistas o conservadores), Minimax: el mínimo de los máximos (para oportunistas), Hurcwiz y Laplace.
- Sin embargo, la mayoría de las veces, los eventos de la naturaleza tienen un grado de incertidumbre, el que puede traducirse en la especificación de una distribución de probabilidad de ocurrencia de los mismos. Esta distribución puede basarse en la historia, antecedentes previos o incluso en las propias creencias del investigador.
- En este caso las decisiones deben incorporar el efecto de la incertidumbre presente en los diferentes eventos de la naturaleza.

2 EL PROBLEMA DE LA INFERENCIA ESTADÍSTICA

Suponemos que un tomador de decisiones (TD), se propone tomar una desición racional bajo condiciones de incertidumbre. Generalmente un TD no tiene certeza del estado de la naturaleza pero puede adquirir información sobre ella a partir de experimentos. Además el TD tiene acciones a realizar optimas dependiendo de los distintos estados a los que fuera enfrentado. (Por ejemplo si usted sale en vehículo de su casa no sabe cómo estará el tránsito pero sabe que si hay mucho atochamiento en la calle tradicional 'X', tomará la alternativa 'A'.

Los elementos de un problema estadístico a ser especificado por el TD para cada problema son los siguientes:

- Espacio de posibles acciones: $A = a$.
- Espacio de los estados posibles de la naturaleza ó espacio de parámetros $\Theta = \theta$.
- Familia de experimentos para adquirir infoemación experimental sobre Θ.
- Espacio muestral, o sea, el espacio de las posibles observaciones experimentales $X = x$.

En pocas palabras la teoría de desiciones sirve para responder a preguntas como las siguientes

- ¿La ley del mineral de una napa subterranea es suficiente para justificar su explotación?
- ¿Una droga para una enfermedad cardíaca es efectiva?
- ¿Una lleva ácida afecta el sistema ecológico?
- ¿Cuál es el diámetro de Plutón?

Hay que tener en cuenta que la manera de responder a estas pregunta depende del tipo de pregunta. Pero todas tienen elementos en común: Cada una requiere obtener datos a partir de una muestra (que también puede ser extraíada con diferentes metodologías).

La Inferencia Estadística trata del problema de sacar conclusiones sobre la población a partir de resultados obtenidos en una muestra.

No olvides notar que, las respuestas a las preguntas anteriores permitirán tomar decisiones. ¿Estas decisiones, que tendrán muchas influencias, son siempre acertadas?. Sabemos que no siempre pues debemos consider la incertidumbre de los resultados debidos al hecho de tomar decisiones a partir de una muestra. En la resolución de este tipo de problemas, se plantea deberá plantear un modelo sobre la pobablación en estudio. En esta guía nos limitaremos a considerar una sola variable X de interés, es decir, X será la medición que se hace sobre los elementos de la población. Como lo vimos en el estudio de la función de distribución empírica, esperamos que la distribución de la muestra sea lo más parecida a la distribución de la población pero lo cierto es que nunca la sabremos con seguridad pues ignoramos si la muestra es realmente *representativa* y no conocemos la distribución de la población. Una manera de proceder consiste en hacer supuestos sobre la función de distribución de la población, lo que constituirá **el modelo estadístico.**

En el caso *paramétrico* se define una familia de funciones de distribución $F(x|\theta)$ para X y el problema se limita a determinar los *parámetros* desconocidos de la distribución $F(x|\theta)$. El conjunto de los valores posibles del parámetro o del vector de parámetros θ es el espacio de parámetros Ω. Por ejemplo:

$F(x\|\theta)$	Ω
$\mathcal{N}(\mu, 1)$	$\mathbb{R}$
$\mathcal{N}(\mu, \sigma)$	$\mathbb{R}\times]0, +\infty[$
$Exp(\beta)$	$]0, +\infty[$
$\mathcal{B}(p)$	$[0, 1]$
$Poisson(\lambda$	$]0, +\infty[$
$Uniforme([\theta_1, \theta_2])$	$\mathbb{R} \times \mathbb{R}$ (sujeto a $\theta_1 < \theta_2$)

Podemos construir a partir de los valores muestrales

- un valor único para θ: es la **Estimación Puntual**
- un intervalo de valores: es la **Estimación por Intervalo**
- elegir entre dos o más conjuntos de valores directamente relacionados con la pregunta inicial: es la **Decisión Estadística**. Un caso particular es la teoría de **Tests Estadísticos de Neyman-Pearson**.

3 TEORÍA DE LA DECISIÓN ESTADÍSTICA

Supongamos que se usted se presentó por primera vez a un concurso de arte en 2003 y fue seleccionado. ¿Cuál es la probabilidad que usted sea seleccionado en el 2004? La falta de más información respecto del concurso podría hacernos concluir que ¡la probabilidad es igual a 1! Si sabemos ahora que solamente el 30% de los concursantes son seleccionados, esta información (que llamaremos *a priori* por ser conocida antes del resultado del concurso) temporizaría su valor de la probabilidad.

Supongamos ahora que nos interesamos ahora en la proporción θ de hombres mayores de más de 45 años con hipertensión. Si no tenemos ninguna otra información relativa al problema diríamos que $\theta \in [0, 1]$, pero que no es 0 ni 1 y tendríamos que escoger un valor al azar, por ejemplo 0.3. Si después de escoger al azar realizamos una encuesta a 8 hombres de más de 45 años y esta arrojó que 5 de 8 hombres son hipertensos. ¿Esto nos permite revisar la estimación anterior? La verdad, no mucho pues la muestra es muy pequeña, nos permitiría a lo más aumentar la primera estimación. Pero, si ahora tomamos una encuesta más extensa, que nos proporcione 350 hombres hipertensos entre 1000, una estimación del orden de 0.35 será más creíble.

En un estudio estadístico hay información inicial a la toma de la muestra (estudios anteriores), que pueden ser muy útiles de agregar a los valores muestrales para así tomar las decisiones o estimar parámetros. Según los supuestos que se hacen sobre los valores de θ tenemos distintos métodos. Cuando se usa información *a priori* sobre θ se habla de **Inferencia Bayesiana**.

3.1 Distribución a priori y distribución a posteriori

Sea $f(x|\theta)$ la función de densidad de X. Si se tiene algunas ideas sobre los valores que puede tomar θ, conviene incorporar este conocimiento **creencia**, en el estudio. Lo anterior se traduce en una distribución de probabilidad sobre el espacio de parámetros Ω, sea $\pi(\theta)$.

Por tanto diremos X se distribuye según $f(X/\theta)$ y θ tiene una distribución a priori $\pi(\theta)$.

Es decir que ahora θ no es un parámetro constante, sino una variable aleatoria. Esta distribución (*priori), no depende de los valores muestrales. Está definida previamente a la toma de la muestra. Si se usa esta distribución* $\pi(\theta)$ *en el problema de inferencia se habla de* **Inferencia Bayesiana**, *en caso contrario se hablará de* **Inferencia Clásica**.

Por ejemplo, en un proceso de fabricación se tiene una proporción θ *desconocida de piezas defectuosas. Si no se sabe nada respecto a* θ, *se puede suponer que todos los valores de* θ *son equiprobables:* $\theta \sim \mathcal{U}(0, 1)$. *Pero, uno puede sopechar que los valores alrededor de 0.10 son más probables; en este caso se podrá tomar una distribución más concentrada en 0.10.*

Definición 3.1 *Se llama* **distribución a priori** *a la distribución atribuida a un parámetro poblacional, antes de tomar alguna muestra.*

Ahora tenemos la función de densidad $f(x|\theta)$ de X y la función de densidad $\pi(\theta)$ del parámetro θ.

Los valores muestrales de una muestra aleatoria de tamaño n constituyen un vector aleatorio $\underline{x} = (x_1, x_2, ..., x_n)$ de componentes pertenecientes todos a la función de densidad $f(x|\theta)$.

Definición 3.2 *Se llama* **función de verosimilitud** *a la densidad conjunta del vector de los valores muestrales; para todo vector observado $\underline{x} = (x_1, x_2, ..., x_n)$ en la muestra, se denota $f_n(\underline{x}|\theta)$.*

Si los valores fueron extraídos de manera independiente, se tiene:

$$f_n(\underline{x}|\theta) = f_n(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

La función de verosimilitud $f_n(\underline{x}|\theta)$ es ahora una densidad condicional y $h(\underline{x}, \theta) = f_n(\underline{x}|\theta)\pi(\theta)$ es la densidad conjunta de $(\underline{x}, \theta)$. De la cual se puede deducir la distribución condicional de θ dado los valores muestrales $\underline{x}$:

Definición 3.3 *La distribución condicional de θ dada la muestra $(x_1, ..., x_n)$ se llama* **distribución a posteriori** *y su densidad es igual a $\xi(\theta|\underline{x}) = \frac{f_n(\underline{x}|\theta)\pi(\theta)}{g_n(\underline{x})}$, en que $g_n(\underline{x}) = \int_\Omega h(\underline{x}, \theta)d\theta$ es la densidad marginal de $\underline{x}$.*

Resumiendo: La variable X que se distribuye según $f(x/\theta)$ y para la cual creíamos que θ se distribuía según una distribución a priori θ, después de la muestra decimos que su parámetro θ tiene una distribución a posteriori (que será utilizada para tomar la decisión), $\xi(\theta|\underline{x})$.

La distribución a posteriori representa la actualización de la información a priori $\pi(\theta)$ en vista de la información contenida en los valores muestrales, $f_n(\underline{x}|\theta)$. Podemos entonces estudiar esta distribución a posteriori de θ dando la moda, la media, la mediana, la varianza, hacer estimaciones puntuales ó estimaciones por intervalo o decidir entre varias hipótesis sobre θ (se verá más adelante).

Ejemplo 1: Sean $X \sim Bernoulli(p)$ y $p \sim \beta eta(\alpha, \beta)$, con α y β dados.

$$f_n(\underline{x}/p) = p^{n\bar{x}_n}(1-p)^{n-n\bar{x}_n}$$

$$\pi(p) = p^{\alpha-1}(1-p)^{\beta-1}/B(\alpha, \beta) \quad 0 \leq p \leq 1$$

en que $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
Sobre el resultado anterior y sabiendo que B es una constante identificamos la densidad a posteriori de p ajustándo sus parámetros, entonces si observamos bien vemos que la distribucióna posteriori de p es:

$$\xi(p|\underline{x}) = p^{\alpha+n\bar{x}_n-1}(1-p)^{\beta+n-n\bar{x}_n-1}/B(\alpha + n\bar{x}_n, \beta + n - n\bar{x}_n)$$

que es la distribución $\beta eta(\alpha + n\bar{x}_n, \beta + n - n\bar{x}_n)$. *Por ejemplo, la moda de esta distribución, cuando está definida, es* $(\alpha - 1)/\alpha + \beta$.

Ejemplo 2: Sean $X \sim \mathcal{N}(\theta, 1)$ *y* $\theta \sim \mathcal{N}(0, 10)$.
$\xi(\theta|\underline{x}) \propto f_n(\underline{x}/\theta)\pi(\theta)$ *(*$\propto$ *se refiere a la proporcionalidad con respecto a* θ*).*

$\xi(\theta/\underline{x}) \propto exp(-\frac{\sum(x_i-\theta)^2}{2} - \frac{\theta^2}{20})$

$\xi(\theta|\underline{x}) \propto exp(-\frac{-11\theta^2}{20} + n\theta\bar{x}_n)$

$\xi(\theta|\underline{x}) \propto exp(-\frac{-11}{20}(\theta - (10n\bar{x}_n/11))^2)$

La distribución a posteriori de θ *es entonces* $\mathcal{N}(\frac{10}{11}n\bar{x}_n, \frac{10}{11})$. *La moda de la distribución es la media* $\frac{10}{11}n\bar{x}_n$.

3.2 Funciones de costo o pérdida

Un aspecto importante del problema de toma de decisiciones estadiísticas es la medición de los errores de decisiones que se cometerían al tomar una desición. Se define para este propósito una función que mide el costo o pérdida de cada decisión.

Consideremos en primer lugar el caso de un conjunto Ω *finito. Este caso puede verse como un juego entre dos jugadores:*

- *Jugador I:* $\Omega = \{\theta_1, \theta_2, .., \theta_r\}$. *Son los valores posibles de* θ *llamados los estados de la naturaleza.*
- *Jugador II:* $\mathcal{D} = \{d_1, d_2, ..., d_p\}$ *que son las decisiones o acciones posibles.*

La función de perdida es la función que mide el costo de tomar una decisón dado un estado de la naturaleza.

Ejemplo: teoría de juego. Consideramos el juego simple en el cual cada jugador muestra 1 o 2 dedos y el jugador I gana si la suma de los dedos es impar y el jugador II gana si la suma es par. El jugador que gana recibe la suma de los dedos mostrados. Se puede considerar el problema del punto de vista del jugador II que tiene que tomar una decisión -mostrar 1 o dos dedos- sin conocer lo que va a jugar el jugador I que es considerado entonces como el estado de la naturaleza -mostrar 1 o 2 dedos-. En la tabla1 se muestra las ganancia y pérdidas del jugador II. El segundo jugador tratará de hacerse una idea de $\pi(\theta = 1) = p$ *la probabilidad a priori con que el jugador I mostrará 1 dedo.*

π	Ω	d_1	d_2
p	$\theta_1 = 1$	-2	+3
1-p	$\theta_2 = 2$	+3	-4

Table 1: Una función de perdida del jugador II

La decisión del jugador II dependerá de p, pero ¿cómo el jugador I decidirá jugar?

Definición 3.4 *Se llama* **función de pérdida** *a la función*
L: $\Omega \times \mathcal{D} \longrightarrow [0, +\infty[$*, en que* $L(\theta, a)$ *mide el costo de tomar la decisión d cuando el estado de la naturaleza (o el parámetro) toma el valor* θ*.*

Dependiendo de como definamos $\mathcal{D}$ *se tienen distintos tipos de métodos de inferencia.*

1. *Si* $\mathcal{D} = \Omega = I\!R$*, se tendrá el problema de estimación puntual de un parámetro real.*
2. *Si* $\mathcal{D}$ *tiene solamente dos puntos* $\{d_1, d_2\}$ *se tendrá el problema de test de hipótesis.*
3. *Si* $\mathcal{D}$ *tiene solamente de un número finito de puntos* $\{d_1, d_2, ..., d_k\}$ *se tendrá el problema de decisiones múltiples.*

3.3 Caso de decisiones múltiples

Consideramos una persona, Don Diego, con m acciones posibles frente a k estados de la naturaleza y una tabla que resuma sus pagos (Tabla 2).

Estados de la naturaleza	Decisiones			
	d_1	d_2	...	d_m
θ_1	L_{11}	L_{12}	...	L_{1m}
θ_2	L_{21}	L_{22}	...	L_{2m}
...	...	...	...	...
θ_k	L_{k1}	L_{k2}	...	L_{km}

Table 2: Tabla de perdida

¿Con que criterio o regla esta persona tomará una decisión?

La respuesta natural es de suponer que Don Diego buscará una acción o decisón que mínimiza su perdida. El problema no tiene una única solución en general. Pero aún si no es tan simple encontrar una solución, es posible eliminar algunas decisiones claramente insatisfactorias.

Definición 3.5 *Si la pérdida* $L(\theta_s, d_j) \leq L(\theta_s, d_i)$ *para todo* $s = 1, 2, \ldots, k$ *y si existe al menos un estado de la naturaleza* s^* *tal que* $L(\theta_s^*, d_j) < L(\theta_s, d_i)$ *(desigualdad estricta), se dice que la decisión* d_j *domina a la decisión* d_i*.*

Toda decisión d_i *dominada se dice* **inadmisible** *y se eliminarán. Las otras decisiones son admisibles.*

Ejemplo En la tabla 3 se observa que la decisión d_2 *es inadmisibel.*

Hay varios criterios posibles para tomar una decisión. Pero hay dos fundamentalmente, que son el criterio del **minimax** *y el criterio de* **Bayes***.*

Estados de la naturaleza	Decisiones d_1	d_2	d_3
θ_1	2	3	1
θ_2	5	10	8
θ_3	10	20	15
θ_4	5	8	7

Table 3: Ejemplo tabla de perdida

3.3.1 Criterio del Minimax

El criterio del **minimax** *está basado en lograr lo mejor de las peores condiciones posibles. Se elige la decisión d_i tal que:*

$$\min_{d_i} \max_{\theta_j}$$

Es un criterio conservador.

Consideremos ahora el proyecto de inversión: Una compañía Comp es duea de unos terrenos en los que puede haber petróleo. Un geólogo consultor ha informado a la gerencia que piensa que existe una probabilidad de $1/4$ *de encontrar petróleo .*

Debido a esta posibilidad, otra compañía petrolera Petro ha ofrecido comprar las tierras en $US90000$*, sin embargo la compañía Comp está considerando conservarla para perforar ella misma. Si encuentra petróleo, la ganancia esperada de la compañía será aproximadamente de* $US70,000$*; incurrirá en una pérdida de* $US100,000$ *si encuentra un pozo seco (sin petróleo).*

Sin embargo, otra opción anterior a tomar una decisión es llevar a cabo una exploración sísmica detallada en el área para obtener una mejor estimación de la probabilidad de encontrar petróleo.

Esta compañía está operando sin mucho capital por lo que una pérdida de $US100,000$ *seria bastante seria.*

De acuerdo al criterio del minimax, la solución del problema es la siguiente:

Estados de la naturaleza	Decisiones Perforar	Vender
petróleo	-700,000	-90,000
seco	100,000	-90,000

Table 4: Tabla de perdida

De esta manera considerando el criterio del minimax, la acción que se debe seguir es vender el terreno de la compañía petrolera y de esta manera la ganancia será de $90,000$*.*

3.3.2 Criterio de Bayes

Estudiaremos aquí el criterio de Bayes y distinguiremos el caso sin datos muestrales (teoría de juegos) y com datos muestrales.

1. *El investigador o jugador conoce solamente las probabilidades $\pi(\theta_1, \ldots, \pi(\theta_k$ y no tiene datos observados, entonces buscará minimizar su perdida esperada. Se calcula para cada decisión d_i:*

$$E(L(\theta, d_j) = \sum_{i=1}^{k} \pi(\theta_i) L(\theta_i, d_j)$$

y se elige la decisión con menor perdida esperada. Es la regla de decisión de Bayes.

Si se aplica este criterio al problema de inversión de la compañía petrolera, suponiendo que la probabilidad de que hay petróleo es 0.25 se tiene lo siguiente:

$$E(L(vender) = -90,000$$
$$E(L(perforar) = -700,000 * 0.25 + 100,000 * 0.75 = -100,000$$

Se concluye que la decisión a tomar es de perforar. Esto muestra lo conservador que es el criterio del minimax.

2. *Se supone ahora que se tienen informaciones adicionales de datos $x_1, x_2, \ldots, x_n$ obtenidos sobre una muestra aleatoria simple. En este caso en vez de utilizar la distribución a priori π para minimizar la perdida esperada (o riesgo), se utilizará la distribución a posteriori*

$$\xi(\theta_i | x_1, \ldots, x_n) = \frac{\pi(\theta_i) f_n(x_1, \ldots, x_n | \theta_i)}{\sum_{t=1}^{k} \pi(\theta_t) f_n(x_1, \ldots, x_n | \theta_t)} \quad i = 1, \ldots, k$$

Definición 3.6 *Una regla de decisión δ es una función que para cada vector de valores muestrales posibles $\underline{x}$ proporciona una decisión $\delta(\underline{x})$ elegida entre las posibles de $\mathcal{D}$.*

Se llama riesgo $\rho_j(\underline{x})$ de seleccionar la decisión d_j cuando se tiene la información $\underline{x}$: $\rho_j(\underline{x}) = E(L(\theta, j) = \sum_{i=1}^{K}$

$$\xi(\theta = \theta_i | \underline{x}) L(_i, _j)$$

Definición 3.7 *Una regla de decisión δ se dice regla de decisón de Bayes si para todo $\underline{x}$, $\delta(\underline{x})$ minimiza el riesgo $\rho(\theta_i | \underline{x})$.*

La ventaja de una regla de decisión de Bayes es de usar más información antes de decidir.

3. *Calculo del riesgo total*

Antes de seleccionar las observaciones, el riesgo que el experimentador afronta por utilizar una regla d edecisión especifica δ se puede calcular de la siguiente manera:

Consideramos A_j el conjunto de las observaciones x para las cuales se elegirá la decisión d_j: $A_j = \{x|\delta(x) = d_j\}$, $\forall j = 1, ..., m$ y supongamos que $f(x|\theta_i)$ es discreta. Si $\theta = \theta_i$:

$$\rho(\delta|\theta = \theta_i) = E(L(\theta_i, \delta(x))|\theta = \theta_i)$$

$$\rho(\delta|\theta = \theta_i) = \sum_{j=1}^{m} L(\theta_i, d_j)I\!P(\delta(x) = d_j|\theta = \theta_i) = \sum_{j=1}^{m} L(\theta_i, d_j) \sum_{x \in A_j} f(x|\theta_i)$$

$$\rho(\delta) = E(\rho(\delta|\theta)) = \sum_{i=1}^{K} \pi(\theta_i)\rho(\delta|\theta = \theta_i) = \sum_{i=1}^{K} \sum_{j=1}^{m} \sum_{x \in A_j} \pi(\theta_i)L(\theta_i, d_j)f(\underline{x}|\theta_i)$$

El riesgo global $\rho(\delta)$ es mínimo cunado la decisión δ es regla de decisión de Bayes.

Nota: Obtener información tiene un costo que debe ser considerado cuando se evalúa el riesgo de una regla de decisión que usa esta información. Es posible que no aporta nada.

Sea $c(\theta_i, x)$ el costo de observar x cuando $\theta = \theta_i$. El costo esperado es:

$$E(c(\theta, x)) = \sum_{\theta} \sum_{x]c(\theta,x)f(x|)\pi(\theta)}$$

Suponiendo la aditividad de los costos, el riesgo total de tomar la decisión δ es entonces: $\rho(\delta) + E(c(\theta, x))$.

La pregunta es entonces Cuánto es el monto $E(c(\theta, x))$ estamos dispuestos a pagar para tener la oportunidad de observar x antes de escoger una decisión?