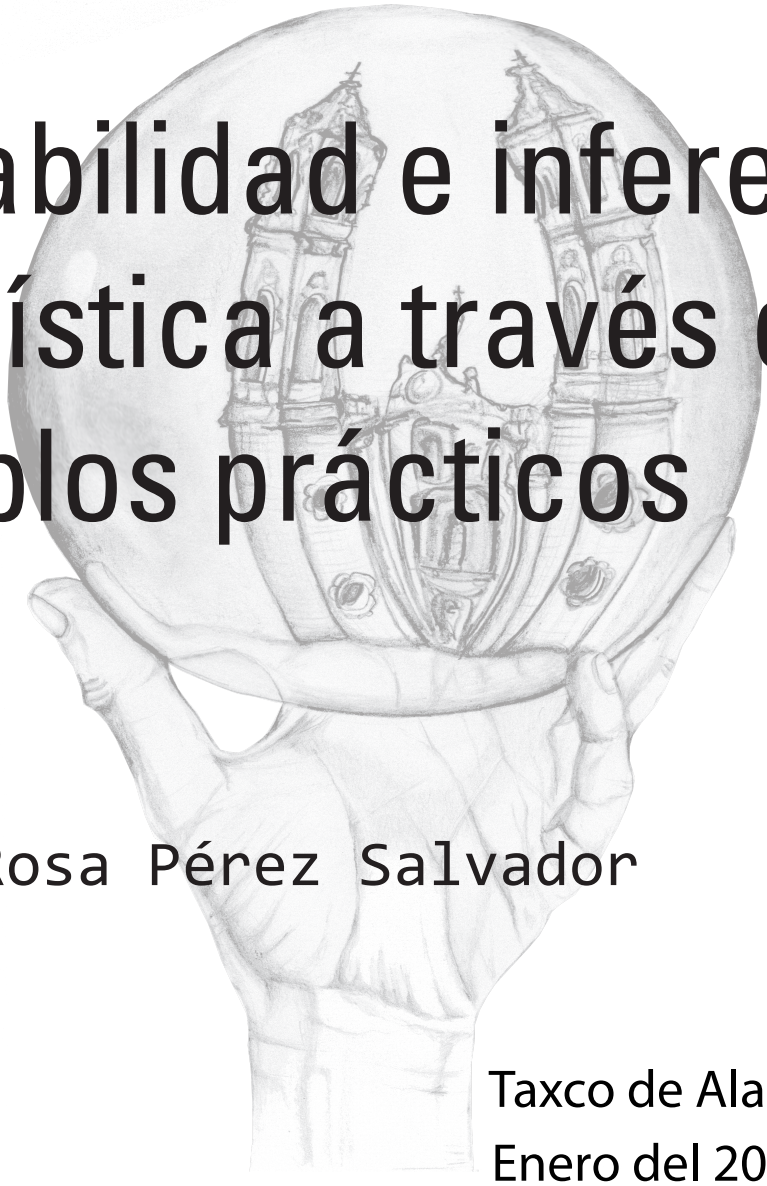
Universidad Autónoma Metropolitana - *Iztapalapa*

UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo

División de Ciencias Básicas e Ingenieria

2° Coloquio

del Departamento de Matemáticas



Probabilidad e inferencia estadística a través de ejemplos prácticos

Blanca Rosa Pérez Salvador

Taxco de Alarcón, Guerrero
Enero del 2009

2do Coloquio del Departamento de Matemáticas

Estadística a través de problemas prácticos

Blanca Rosa Pérez Salvador

Comité Organizador

Mario Pineda Ruelas
Roberto Quezada Batalla
Blanca Rosa Pérez Salvador
Luis Aguirre Castillo
Daniel Espinosa
Constancio Hernández García
Michaell Rivera Arce (Apoyo logístico)

Material didáctico del Departamento de Matemáticas de la Universidad Autonóma Metropolitana-Iztapalapa con el apoyo del Programa Integral de Fortalecimiento Institucional 3.3 (PIFI).

Estadística a través de problemas prácticos

Blanca Rosa Pérez Salvador

Departamento de Matemáticas, UAM-I

Casa abierta al tiempo

Universidad Autónoma Metropolitana

Contenido

Capítulo 1

Introducción

La teoría de la Estadística es una de las ramas de las matemáticas que se utiliza para resolver problemas en casí todas las actividades de la vida humana. Por lo tanto con la estadística se puede resolver un sin número de problemas prácticos.

La estadística, se divide en dos grandes ramas: la **estadística descriptiva** y la **estadística inferencial** cuyas definiciones formales son:

DEFINICIÓN 1.1. La *estadistica descriptiva* consiste en el conjunto de métodos que permiten clasificar, ordenar y presentar en gráficas o tablas a los datos de una muestra.

Estos métodos nos permiten tener una una descripción de como están distribuidos los datos de una muestra.

DEFINICIÓN 1.2. La *estadística inferencial* consiste en el conjunto de métodos que se utilizan para obtener información acerca de los parámetros de una población con únicamente información parcial de la misma.

Este trabajo tratará unicamente sobre algunos aspectos de la estadística inferencial que es la más interesante y la de mayor relevancia en las aplicaciones, Todos los métodos y técnicas que conforman la estadistica inferencial se clasifican en dos grupos, estimación de parámetros y prueba estadística de hipótesis. La estadística inferencial se auxilia de la teoría de las probabilidades, debido a que los datos de la muestra utilizada se obtienen por un proceso aleatorio, y al ser la muestra aleatoria se requiere conocer algunos conceptos relacionados con la probabilidad. Por lo tanto para iniciar el estudio se la estadísitica se requiere conocer algunos conceptos de la teoría de la probabilidad, que es el material con el que iniciaremos; se darán las definiciones de algunos conceptos probabilisticos, sin entrar al estudio formal de la probabilidad.

1.1. Breve introducción a la probabilidad

La probabilidad es una rama de las matemáticas que se encarga del estudio de los modelos frecuentistas. En probabilidad se habla de experimentos realizados, o experimentos por realizar. la definición de experimento en probabilidad difiere del concepto de experimento en física o química. La definición de experimento en probabilidad se da enseguida.

DEFINICIÓN 1.3. Un *experimento* en la teoría de la probabilidad es algo suceptible de ser observado y ser repetido.

Algunos ejemplos de experimentos son: 1) Lanzar un volado. 2) Lanzar un dado. 3) Observar si llueve o no llueve. 4) Medir el tiempo de espera del próximo autobús. 5) Observar el número de accidentes automovilisticos en un año.

DEFINICIÓN 1.4. El *espacio muestral* es el conjunto de posibles resultados al realizar un experimento. Al espacio muestral se le denota con la letra Ω

Por ejemplo, si el experimento es lanzar un dado entonces el espacio muestral correspondiente es $\Omega = \{1, 2, 3, 4, 5, 6\}$; si el experimento es lanzar una moneda, el espacio muestral es $\Omega = \{\acute{a}guila,\ sol\}$, si el experimento es ver un nacimiento, el espacio muestral es $\Omega = \{$niño, niña$\}$, etc.

DEFINICIÓN 1.5. Un *evento* es un subconjunto del espacio muestral.

Por ejemplo, si el experimento es lanzar un dado, dos posibles eventos son $\{1, 2\}$ y $\{2, 4, 6\}$

Algunos eventos particulares son:

Eventos elementales: son eventos que no pueden subdividirse en eventos más pequeños, por ejemplo en el lanzamiento de un dado un evento elemental es $E = \{4\}$, y el eveno $F = \{2, 5\}$ no es elemental porque puede descomponerse en los eventos $\{2\}$ y $\{5\}$, estos últimos ya son eventos elementales.

Evento imposible: es el evento que no puede ocurrir y se denota con el símbolo del conjunto vacio, ϕ. Por ejemplo es imposible que al lanzar una moneda resulte que cae de canto, o que la moneda desaparezca en el aire.

Evento seguro: es el evento que siempre ocurre, y corresponde al espacio muestral. Por ejemplo, en una rifa, se juega con el evento seguro si se adquieren todos los boletos de la misma.

Eventos mutuamente excluyentes: Son eventos que no pueden ocurrir simultaneamente. La ocurrencia de uno de ellos excluye

la posibilidad de ocurrencia del otro. A y B son mutuamente excluyentes si y sólo si $A \cap B = \phi$. Por ejemplo, al lanzar un dado, la aparición del número 2 excluye la aparición del n'umero 4, es imposible que salgan los dos simultáneamente.

DEFINICIÓN 1.6. La *probabilidad* es una medida de la oportunidad que ocurra un evento y se denota con la letra P; así la probabilidad del evento A se escribe como $P(A)$

Axiomas de probabilidad. Son tres las propiedades básicas de la probabilidad.

(1) $P(A) \geq 0$ para cada evento A
(2) $P(\Omega) = 1$
(3) $P(A \cup B) = P(A) + P(B)$ si $A \cap B = \phi$

A partir de estos tres axiomas se pueden demostrar los siguientes teoremas.

TEOREMA 1.7. *Si A es un evento, entonces $P(A^c) = 1 - P(A)$*

DEMOSTRACIÓN. Se sabe que

- $A \cup A^c = \Omega$ y
- $A \cap A^c = \phi$,

Por un lado $P(A \cup A^c) = P(\Omega) = 1$, por otro lado, el axioma 3 afirma que $P(A \cup A^c) = P(A) + P(A^c)$, de estas dos ecuaciones se sigue que

$$P(A) + P(A^c) = 1 \quad \Rightarrow \quad P(A^c) = 1 - P(A) \qquad \square$$

TEOREMA 1.8. $P(\phi) = 0$

DEMOSTRACIÓN. Por el teorema anterior, se sigue que

$$P(\phi) = P(\Omega^c) = 1 - P(\Omega) = 1 - 1 = 0 \qquad \square$$

TEOREMA 1.9. *Si A y B son dos eventos tales que $A \subset B$ entonces $P(A) \leq P(B)$.*

DEMOSTRACIÓN. El evento B se puede escribir como la unión de dos eventos mutuamente excluyentes, $B = A \cup (B \cap A^c)$. Por el axioma 3, se sigue que $P(B) = P(A) + P(B \cap A^c) \geq P(A)$, ya que por el axioma 1 se satsface que $P(B \cap A^c) \geq 0$. $\square$

TEOREMA 1.10. *Si A y B son eventos, entonces*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

DEMOSTRACIÓN. Se tiene que $A \cup B = A \cup (A^c \cap B)$ y $A \cap (A^c \cap B) = \phi$ entonces por el tercer axioma de probabilidad se sigue que $P(A \cup B) = P(A) + P(A^c \cap B)$.

Por otra parte, se tiene que $B = (A \cap B) \cup (A^c \cap B)$ y $(A \cap B) \cap (A^c \cap B) = \phi$ entonces $P(B) = P(A \cap B) + P(A^c \cap B)$, de donde se sique que $P(A^c \cap B) = P(B) - P(A \cap B)$, combinando estos dos resultados se llega a

$$P(A \cup B) = P(A) + P(A^c \cap B) = P(A) + P(B) - P(A \cap B) \qquad \square$$

Intuitivamente, la probabilidad corresponde a la frecuencia de ocurrencia de un resultado en un experimento. Uno de los casos más simples de calcular la probabilidad es cuando el espacio muestral es finito y todos sus elementos tienen la misma oportunidad de ocurrir. A estos espacios muestrales de les conoce como equiprobables y si $A \subset \Omega$ la probabilidad de A se define como

$$P(A) = \frac{\#A}{\#\Omega},$$

que es la definición clásica de probabilidad.

1.2. Variables aleatorias

DEFINICIÓN 1.11. Una *variable aleatoria* es una función cuyo dominio es el espacio muestral y cuyo codominio son los números reales; una variable aleatoria es de la forma

$$X \colon \Omega \to \mathbb{R}.$$

Las variables aleatorias se denotan con las últimas letras mayúsculas del alfabeto, esto es X, Y y Z.

EJEMPLO 1.12. Se lanzan tres monedas al aire, el espacio muestral es

$$\Omega = \{(sss),\ (ssa),\ (sas),\ (ass),\ (saa),\ (asa),\ (aas)\}$$

Considere la variable aleatoria

$$X \colon \Omega \to \mathbb{R}$$

tal que X = número de águilas. La función es:

$$X(sss) = 0 \qquad X(ssa) = 1 \qquad X(sas) = 1 \qquad X(ass) = 1$$
$$X(saa) = 2 \qquad X(asa) = 2 \qquad X(aas) = 2 \qquad X(aaa) = 3$$

DEFINICIÓN 1.13. Una variable aleatoria cuya imagen es un número finito o infinito numerable se conoce como *variable aleatoria discreta.*

DEFINICIÓN 1.14. La función de probabilidad, o función de densidad de una variable aleatoria es igual a

$$f(x) = P(X = x)$$

EJEMPLO 1.15. La función de probabilidad de la variable aleatroria que determina el número de águilas es igual a

$f(0) = 1/8$
$f(1) = 3/8$
$f(2) = 3/8$
$f(3) = 1/8$

La función de probabilidad de una variable aleatoria satisface las siguientes propiedades:

- $f(x) \geq 0$
- $\sum_x f(x) = 1$

La media de una variable aleatoria es una medida de posición, la fórmula para calcular la media es

$$\mu = E(X) = \sum_x x f(x)$$

La media de la variable que indica el número de águilas al lanzar tres monedas es

$$E(X) = 0f(0) + 1f(1) + 2f(2) + 3f(3) = 1(3/8) + 2(3/8) + 3(1/8) = 12/8 = 1.5$$

$\mu = 1.5$ indica la posición promedio de los valores $0,\ 1,\ 2,\ 3$

La varianza de una variable aleatoria es una medida de la dispersión de los datos. Se calcula con la fórmula

$$V(X) = \sum_x (x - \mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2$$

La varianza de la variable que indica el número de águilas al lanzar tres monedas es

$$\begin{aligned} V(X) &= (0 - 1.5)^2 f(0) + (1 - 1.5)^2 f(1) + (2 - 1.5)^2 f(2) + (3 - 1.5)^2 f(3) \\ &= 2.25(1/8) + 0.25(3/8) + 0.25(3/8) + 2.25(1/8) = 0.75 \end{aligned}$$

Ejercicios

1. Enliste los elementos de los espacios muestrales de los siguientes experimentos:
   - Se lanzan dos monedas al aire.
   - En una caja hay dos canicas rojas y una canica blanca, se seleccionan sin reemplazo dos canicas de la caja.
   - En la caseta de cobro de una autopista se cuenta el número autos que pasan durante diez minutos.
   - Una compañia de seguros tiene 1000 automoviles asegurados por da;os para el pr'oximo año.

2. En una caja se tienen 5 canicas, dos rojas y tres verdes, se eligen al azar sin reemplazo dos canicas de la caja, (a) enliste los elementos del espacio muestral. (b) Encuentre la probabilidad de que las dos bolas sean verdes. (c) Encuentre la probabilidad de que las dos bolas sean rojas. (d) Encuentre la probabilidad de que una bola sea roja y la otra sea verde.
3. Se elije una persona al azar de un grupo de 10 mujeres y 15 hombres. 3 de las mujeres fuman y 7 de los hombres fuman. Sean los eventos A: la persona seleccionada es mujer, y B: la persona seleccionada fuma. Encuentre la probabilidad de que $P(A \cup B)$.
4. Sean A, B y C eventos, demuestre que

$$P(A\cup B\cup C) = P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$$

5. En un servicio de autolavado pueden llegar 0, 1, 2, 3, 4, o 5 automoviles con probabilidad igual a

   $P(X = 0) = 0.05$
   $P(X = 1) = 0.10$
   $P(X = 2) = 0.40$
   $P(X = 3) = 0.20$
   $P(X = 4) = 0.15$
   $P(X = 5) = 0.10$

   Encuentre la media y la varianza del número de autos que llegan a solicitar servicio en un día.

Capítulo 2

Estimación de parámetros

Como ya se dijo, las técnicas de inferencia estadística caen en una de dos posibles categorías: **estimación de parámetros** y **pruebas de hipótesis estadísticas**. En este capítulo se revisarán dos problemas en los que se introduciran los elementos de la estimación de parámetros.

Un parámetro es un valor relacionado con la población objetivo y por lo regular es desconocido. Por ejemplo, el promedio del ingreso en la población, el número de elementos en la población, etc.

Un estimador del parámetro es una función de los valores muestrales, con el cual pretendemos conocer el valor del parámetro. Por ejemplo, el promedio de una muestra estima el valor de la media de la población.

La estadística nos proporciona métodos para obtener estimadores y criterios para establecer las propiedades deseables en los estimadores.

Problema 1.
¿Cuántos Peces Hay en el Lago?

El guardabosques de un parque nacional quiere saber cuántas carpas adultas habitan en el lago del parque, pues debe decidir si se permite la pesca deportiva o si se declara un estado de veda.

El guardia llega a la conclusión que no es posible contar todas las carpas directamente; por lo que debe idear la manera de obtener indirectamente una estimación aceptable de N el total de carpas adultas en el lago. El guardia considera que en el lago hay N carpas adultas. N es un número asociado a la población, y por lo tanto es el **parámetro poblacional** y es desconocido.

La distribución uniforme de los peces en el lago. Después de cavilar un rato, el guardia considera haber encontrado una posible solución; como se tienen los medios técnicos para determinar con suficiente precisión el volumen total del lago V (por ejemplo con ultrasonido) piensa que si delimita una pequeña región en el lago, de

volumen v, y cuenta los peces que hay en ese espacio, entonces puede utilizar esta información para obtener un estimador de N.

Entonces se tendrían los siguientes valores,

- N es el número de carpas adultas en el lago;
- V es el volumen del lago;
- v es el volumen en la región controlada por el guardia;
- n es el número de carpas adultas en el volumen v.

El guardia forestal considera que el número de peces en la región pequeña controlada, es proporcional al número de peces en el volumen total del lago.

Bajo este supuesto de proporcionalidad, se satisface la relación:

$$\frac{N}{V} = \frac{n}{v}$$

de donde se puede despejar el valor de N

$$N = \frac{n}{v}V$$

Una vez modelado el problema y con una fórmula para resolverlo cabe preguntarse:

¿Qué tan válido es este valor como estimador de N?. Para deducir esta fórmula, se considero que los peces se encuentran uniformemente distribuidos en todo el lago. Pues si el número de peces en el lago es proporcional al número de peces en la región controlada es porque en todo el lago están los peces distribuidos uniformemente, como se observa en la figura (a).

(a) Distribución uniforme de los peces en todo el lago	(b) Distribución de los peces en forma de cardúmenes

Sin embargo; esta supocisión no es cierta, ya que los peces se encuentran reunidos en cardúmenes y no diseminados uniformemente por todo el lago como se observa en la figura (b).

De ahí que el supuesto que el número de carpas en la región controlada, sea proporcional al número total de peces en el lago es falso y entonces la ecuación de proporcionalidad no corresponde a la realidad.

Además la presencia del guardia puede alterar la conducta de los peces, los peces pueden huir al ver al guardia, o pueden reaccionar acercándose al mismo. y puede ser difícil contarlos dentro de la región controlada de volumen v.

Esto significa que al contrastar la solución propuesta con la realidad los resultados son insatisfactorios. De manera que ni es fácil tomar la muestra, ni la suposición del modelo es acorde con la realidad.

Es decir: el modelo no es una buena aproximación con la realidad y resulta inadecuado.

El guardia necesita proponer otro modelo que sea acorde con la conducta de los peces.

Método de captura y recaptura. Este método de solución tiene dos etapas:

- **Primera Etapa**: se sacan del lago M carpas adultas, se les marca con una señal y se regresan vivas al lago; después de esto, se deja un tiempo razonable para que las carpas marcadas se integren al resto de las carpas.

  Al final de esta primera etapa, en el lago existen dos clases de carpas, las carpas marcadas y las carpas sin marcar.

- **Segunda Etapa**: se extraen al azar n carpas adultas del lago (a este subconjunto de carpas se denomina "la muestra de tamaño n", y luego se cuenta el número de carpas marcadas que contiene esta muestra. Suponga que en la muestra de n carpas se observan m carpas marcadas.

Se tiene entonces que:

- N es el número total de carpas adultas en el lago:
- M es el número total de carpas marcadas en el lago.
- n es el número total de carpas en la muestra.
- m es el número de carpas marcadas en la muestra.

Bajo el supuesto que las carpas marcadas y las no marcadas se encuentran mezcladas uniformemente en el cardumen, se puede suponer que el número de carpas marcadas en la muestra es proporcional al número de carpas marcadas en la población total, esto es, se debe satisfacer la relación:

$$\frac{N}{M} = \frac{n}{m} \tag{2.1}$$

de donde al despejar el valor de N, se obtiene

$$N = \frac{n}{m}M \tag{2.2}$$

Igual que antes, cabe preguntarse sobre que tan bueno es este número como estimador del parámetro N.

Al establecer la relación 4.1, se consideró que los peces marcados se encuentran uniformemente distribuidos con el resto de la población total.

Este supuesto a nuestro entender se cumple razonablemente, pues no es creible que los peces marcados estén todos juntos formando un grupo sin mezclarse con el resto de la población de peces; por lo que consideramos que el modelo propuesto parece ser adecuado para estimar a N.

Pero, a pesar de ser éste un modelo razonablemente adecuado, el valor obtenido con la fórmula 4.2 no corresponde al verdadero valor de N, porque en el proceso del muestreo interviene el azar, esto es, el número de peces marcados en la muestra es una variable aleatoria, y puede cambiar si se otra muestra de tamaño n en circunstancias semejantes.

Para distinguir el estimador del valor real (o parámetro), al estimador de N y se denota con la misma letra N pero con un gorrito arriba.

$$\hat{N} = \frac{n}{m}M \tag{2.3}$$

N es el parámetro poblacional y $\hat{N}$ es el estimador de N.

Como ejemplo se va a revisar el caso en que en total se tiene 20 peces y que de estos, 4 están marcados; es decir, N=20 y M=4.

Si se elijen 10 de estos peces al azar, en la muestra puede haber desde cero a cuatro peces marcados, y únicamente cuando en la muestra hay 2 peces marcados es que se satisface satisface la relación 4.1,

$$\frac{20}{4} = \frac{10}{2}$$

y por lo tanto, la estimación coincide con el del parámetro, esto es $\hat{N} = 4 \times 10/2 = 20$.

El número de peces marcados en la muestra, X, es un número aleatorio que puede ir de cero a cuatro ($X = 0,\ 1,\ 2,\ 3,\ 4$) por lo tanto los valores que puede tomar el estimador $\hat{N} = 4 \times 10/x$ son los siguientes:

- Cuando $X = 0$ se tiene que $\hat{N} = \infty$
- Cuando $X = 1$ se tiene que $\hat{N} = 40$
- Cuando $X = 2$ se tiene que $\hat{N} = 20$
- Cuando $X = 3$ se tiene que $\hat{N} = 13.33$
- Cuando $X = 4$ se tiene que $\hat{N} = 10$

El estimador de N es una variable aleatoria y podemos preguntarnos con que probabilidad puede tomar los diferentes valores posibles.

La probabilidad se puede calcular usando la definición clásica.

$$P(X = m) = \frac{Número\ total\ de\ muestras\ de\ tamaño\ n\ con\ m\ peces\ marcados}{Número\ total\ de\ muestras\ de\ tamaño\ n}$$

De esta manera nos encontramos ante el problema de tener que contar, lo cual nos lleva a recordar el cálculo combinatorio, en particular la definición de combinaciones.

Número total de muestras de tamaño n (número de elementos del espacio muestral). Observe que la muestra es un subconjunto de n elementos del conjunto total que tiene N elementos, entonces para saber cuántas muestras de tamaño n es posible obtener, se utiliza la fórmula de las combinaciones, esto es

$$\text{El total de muestras de tamaño } n = \begin{pmatrix} N \\ n \end{pmatrix}$$

Ahora se va a obtener una fórmula para determinar cuántas de estas posibles muestras tienen m peces marcados.

Número total de muestras de tamaño n con m peces marcados (Casos favorables). Del total de posibles muestras de tamaño n, interesa determinar cuántas de ellas tienen exactamente m carpas marcadas, para ello se debe ver que las m carpas marcadas en la muestra, es un subconjunto de las M carpas marcadas en el lago, y que las $n - m$ carpas no marcadas en la muestra, es un subconjunto de las $N - M$ carpas no marcadas en el lago, de aquí se sigue que:

El total de subconjuntos de carpas marcadas de las M que hay en el lago es: $\begin{pmatrix} M \\ m \end{pmatrix}$

El total de subconjuntos de $n - m$ carpas no marcadas de las $N - M$ que hay en el lago es: $\begin{pmatrix} N - M \\ n - m \end{pmatrix}$

Por la regla del producto de los métodos de conteo, se sigue que el total de muestras de tamaño n con exactamente m carpas marcadas es el producto de estos dos números, esto es:

$$\text{El total de muestras de tamaño } n \text{ con } m \text{ carpas marcadas} = \begin{pmatrix} M \\ m \end{pmatrix} \begin{pmatrix} N - M \\ n - m \end{pmatrix}$$

entonces la probabilidad de tener una muestra con exactamente $X = m$ carpas marcadas es igual al cociente:

$$P(X = m) = \frac{\begin{pmatrix} M \\ m \end{pmatrix}\begin{pmatrix} N-M \\ n-m \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

Ahora se presentarán cuatro ejemplos para explorar el comportamiento de esta fórmula en función de la variable m.

Ejemplo 1

Considere que

$N = 20$

$M = 8$ y

$n = 6$

Si en la segunda etapa de captura se eligen 6 peces, El valor de m (el número de peces marcados en la muestra) puede tomar los valores de 0 a 6.

En cada caso se tendría un valor del estimador con una cierta probabilidad. En la siguiente tabla se muestra los valores de m, los valores de $\hat{N}$ y las respectivas probabilidades,

m	$\hat{N}$	Probabilidad
0	∞	0.02383901
1	48	0.16346749
2	24	0.35758514
3	16	0.31785346
4	12	0.11919505
5	9.6	0.01733746
6	8	0.00072239

En la tabla se observa que los valores del estimador que se encuentran más cerca a $N = 20$, son 16 y 24. La altura de las barras de la gráfica corresponden a la probabilidad de cada valor. Se observa que los valores 16 y 24 son los mas probables y juntos tienen una probabilidad igual a 0.675438. También se puede observar que los posibles valores del estimador tienen una gran variabilidad, pues va de 8 a ∞.

Ejemplo 2

$N = 20$,

$M = 8$ y

$n = 5$.

En este caso, el rango de X es de 0 a 5 ($X = 0,\ 1,\ 2,\ 3,\ 4$ y 5) y cuando $X = 2$ se satisface la ecuación 4.1 por lo que la estimación coincide con el verdadero valor de $N = 20$. En la siguiente tabla se presentan los valores del estimador y sus probabilidades respectivas,

m	$\hat{N}$	Probabilidad
0	∞	0.05108359
1	40	0.25541796
2	20	0.39731682
3	13.333	0.23839009

Observe que la barra asociada al número 20 es la más alta, lo que indica que este valor es el más probable.

También se puede observar que los posibles valores del estimador tienen una gran variabilidad, pues va de 8 a ∞.

Ejemplo 3
$N = 100$,
$M = 20$ y
$n = 10$.

El valor de m (el número de peces marcados en la muestra) puede tomar los valores de 0 a 10. En cada caso se tendría un valor del estimador con una cierta probabilidad. En la siguiente tabla se muestra los valores de m, los valores de $\hat{N}$ y las respectivas probabilidades,

□

m	$\hat{N}$	Probabilidad
0	∞	0.09511627
1	200	0.26793316
2	100	0.31817063
3	66.667	0.20920809
4	50	0.0841073
5	40	0.02153147
6	33.333	0.00354136
7	28.571	0.00036793
8	25	2.2996E-05
9	22.222	7.7623E-07
10	20	1.0673E-08

El mejor valor para estimar a $N = 100$ es cuando $m = 2$, porque en este caso $\hat{N} = 100$. También en este caso, es el valor más probable. La variación de los valores del estimador tiene una variación muy grande, va del 20 al infinito.

Ejemplo 4
$N = 500$,
$M = 100$ y
$n = 50$.

El rango de X es 0 a 50 ($X = 0, 1, 2, 3, \ldots, 50$) y una tabla con 51 entradas es muy grande para escribirla aquí, por lo que sólo se presenta la gráfica con las probabilidades de los valores que puede tomar el estimador.

□

También en este ejemplo $\hat{N} = 500$ es el valor más probable y el estimador tiene gran variación.

En los cuatro ejemplos revisados se observó que el valor del estimador más cercano al verdadero valor del parámetro es el más probable; sin embargo, los posibles valores de la estimación presentan una gran variación, por lo que el estimador podría estar muy lejos del parámetro que estiman.

Método de captura y recaptura modificado. El estimador $\hat{N} = nM/m$ tiene el inconveniente de que se requiere dividir entre 0, cuando $m = 0$, esto provoca que su dispersión sea muy grande.

Este inconveniente se puede eliminar si modificamos el esquema de muestreo.

Este nuevo método tiene también dos etapas, la primera es exactamente igual que la del método anterior, para tener peces marcados y no marcados en el lago.

- **Primera Etapa**: se sacan del lago M carpas adultas, se les marca con una señal y se regresan vivas al lago; después de esto, se deja un tiempo razonable para que las carpas marcadas se integren al resto de las carpas.
- **Segunda Etapa**: se extraen al azar sucesivamente una carpa tras otra y se detiene el procedimiento cuando se obtengan m carpas marcadas (m un número preestablecido).

En este caso el total de observaciones es aleatorio. Si el número de extracciones es n significa que con la extracción $n-1$ ya se tienen en la muestra $m-1$ peces marcados, y que en la extracción siguiente se obtiene el último pez marcado con el que se completa los m peces en la muestra.

□

En este punto se tiene que:

- N es el número total de carpas adultas en el lago:
- M es el número total de carpas marcadas en el lago.
- n es el número total de carpas extraidas.
- m es el número de carpas marcadas en la muestra.

El número n es aleatorio y m es fijo, pues de antemano se indica cuántos peces marcados se quieren en la muestra y no se sabe apriori el número de extracciones necesarias para seleccionar exactamente m carpas marcadas.

La proporción que se propone en este caso es

$$\frac{N-M}{M+1} = \frac{n-m}{m}$$

y al despejar el valor de N se obtiene un nuevo estimador

$$\hat{N} = \frac{(n-m)(M+1)}{m} + M$$

Los valores $M+1$ y m son constantes conocidas en esta fórmula. Cuando m divide a $M+1$ los valores del estimador son todos enteros, si esto no ocurre, los valores del estimador pueden no ser enteros. Ahora se analizara la probabilidad de este estimador.

Casos totales. La extracción de las carpas se realiza sucesivamente hasta tener m carpas marcadas y para determinar la cardinalidad del espacio muestral se consideran dos etapas: la primera es cuando se

obtienen las primeras $n-1$ extracciones, la segunda etapa es cuando se obtiene la última carpa que debe estar marcada.

Los primeros $n-1$ peces seleccionados forman un subconjunto de la población total de carpas, y el total de formas en que se pueden elegir es igual a $\binom{N}{n-1}$ y el último pez seleccionado puede ser cualquiera de los $N-n+1$ que quedan en el lago. Por la regla del producto de los métodos de conteo, se tiene que el total de formas de extraer n carpas del total de carpas, siendo la última una carpa marcada es:

$$\binom{N}{n-1}(N-n+1)$$

Casos Favorables. Del total de posibles muestras de tamaño n interesa determinar cuántas tienen exactamente $m-1$ carpas marcadas en las $n-1$ primeras extracciones.

Las $m-1$ carpas marcadas es un subconjunto de las M carpas marcadas, y las $n-m$ carpas no marcadas es un subconjunto de las $N-M$ carpas no marcadas.

El total de subconjuntos de $m-1$ carpas marcadas de las M carpas marcadas es igual a $\binom{M}{m-1}$

El total de subconjuntos de $n-m$ carpas no marcadas del total de $N-M$ carpas no marcadas es igual a $\binom{N-M}{n-m}$

La última carpa marcada puede ser cualquiera de las $M-m+1$ carpas marcadas que permanecen aún en el lago. Por la regla del producto de los métodos de conteo, el total de muestras con exactamente m carpas marcadas es el producto de estos tres terminos, esto es:

El total de muestras de tamaño n con m carpas marcadas es

$$\binom{M}{m-1}\binom{N-M}{n-m}(M-m+1)$$

De donde se tiene que la probabilidad de detenerse en la extracción n, $(X=n)$, es igual al cociente

$$P(X=n)=\frac{\binom{M}{m-1}\binom{N-M}{n-m}(M-m+1)}{\binom{N}{n-1}(N-n+1)}=\frac{m\binom{M}{m}\binom{N-M}{n-m}}{n\binom{N}{n}}$$

Para estudiar esta fórmula es conveniente verla gráficamente para algunos casos particulares.

n	Estimador de N	Probabilidad
3	8	0.049122807
4	11	0.104024768
5	14	0.143034056
6	17	0.158926729
7	20	0.153250774
8	23	0.132031436
9	26	0.102691117
10	29	0.072017147
11	32	0.045010717
12	35	0.024430266
13	38	0.01100262
14	41	0.00371517
15	44	0.000722394

Ejemplo 5

Considere que

$N = 207$,

$M = 8$ y

$m = 3$ (con esto se tiene que $(M+1)/m$ es un entero

En este caso los posibles valores que puede tomar la variable n es de 3 a 15 ($n = 3,\ 4,\ 5,\ \ldots,\ 15$); en la siguiente tabla se presentan los valores del estimador y sus probabilidades,

□

Se puede observar que el mejor valor de $\hat{N}$ es cuando $\hat{N} = 20$, este valor tiene una probabilidad alta de ocurrir. Además se puede ver que este estimador tiene menor variación que con el esquema original de muestreo de captura y recaptura.

Ejemplo 6

Considere que

$N = 100$,

$M = 20$ y

$m = 3$ (con esto se tiene que $(M+1)/m$ es un entero

En este caso los posibles valores que puede tomar la variable n es de 3 a 83 ($n = 3,\ 4,\ 5,\ \ldots,\ 83$); en la siguiente gráfica se presentan los valores del estimador y sus probabilidades.

□

Se observa que alrededor de 100 se encuentran los valores más probables.

Conclusion: Los dos esquemas de muestreo proporcionan estimadores de N cuyos valores más cercanos al valor real de N son más probables, sin embargo, el método de captura y recaptura modificado proporciona estimadores con menos dispersión.

Problema 2
¿Cuántos tanques tiene el enemigo?

Ahora se va a estudiar un problema diferente. Se considera dos ejercitos que están en guerra y los estrategas de uno de los ejercitos quieren determinar el número de tanques que tiene el otro ejercito.

□

Es claro que se puede conocer este número revisando los inventarios del otro ejercito, pero esto no es posible y por lo tanto se debe recurrir a un método indirecto.

□

Se ha visto que los tanques observados del otro ejercito traen escrito un número, lo que sugiere que todos los tanques deben estar numerados del 1 al N, siendo N el total de tanques.

□

Además, se puede considerar que es igualmente probable observar a cualquiera de los tanques del enemigo. El reto es entonces estimar el valor de N.

□

De esta manera se observa el número que traen escrito 5 tanques que fueron observados en diferentes lugares, los números observados son: 81, 47, 16, 97, 58. Estos datos son nuestra muestra aleatoria.

□

Con estos cinco números ?'qué podemos decir del valor de N?

Para tener diferentes opciones, la tarea de determinar el valor de N se la encomiendan a tres grupos diferentes que trabajaran independientemente.

Para estimar el número N, primero se ordenan los datos de la muestra.

$$16,\ 47,\ 58,\ 81,\ 97$$

Propuesta de estimación del equipo 1

El primer equipo de estrategas considera que si los tanques están numerados sucesivamente entonces N debe ser mayor o igual a 97. Y entonces proponen como estimador, precisamente a este número.

$$\hat{N}_1 = \max\{X_1,\ X_2,\ X_3,\ X_4,\ X_5\} = \max = 97$$

El primer equipo estima que el ejercito contrario tiene 97 tanques.

Propuesta de estimación del equipo 2

El segundo equipo de estrategas también considera que si los tanques están numerados sucesivamente entonces N debe ser mayor o igual a 97, pero piensa que si el tanque número uno no fue observado, bien podría ser que tampoco se hubiera observado el último tanque (el marcado con el número N), y para describir esto escriben así nuevamente la lista de datos:

$$1,\ 16,\ 47,\ 58,\ 81,\ 97,\ N$$

Este equipo considera que la separación entre max = 97 y N debe ser similar a la separación entre min = 16 y 1, esto es

$$N - \max = \min - 1 \ \Rightarrow\ N = \max + \min - 1$$

El estimador propuesto por el segundo equipo de estrategas es

$$\hat{N}_2 = \max + \min - 1 = 97 + 16 - 1 = 112$$

Propuesta de estimación del equipo 3

El tercer equipo de estrategas también considera que si los tanques están numerados sucesivamente entonces N debe ser mayor o igual a 97, y razonan que las separaciones entre dos números observados debe ser semejante, por lo que calculan el promedio de las separaciones entre dos datos consecutivos.

$$1, \quad 16, \quad 47, \quad 58, \quad 81, \quad 97, \quad N$$

$$d_1 \quad d_2 \quad d_3 \quad d_4 \quad d_5 \quad d$$

Proponen que

$$d = \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5}$$

$$= \frac{(16-1) + (47-16) + (58-47) + (81-58) + (97-81)}{5} = \frac{97-1}{5}$$

Esto lleva a la ecuación

$$N - \max = d \quad \Rightarrow \quad N = \max + d = \frac{6\max - 1}{5}$$

Y el estimador propuesto por el tercer equipo es

$$\hat{N}_3 = \frac{(n+1)\max - 1}{n} = \frac{6 \times 97 - 1}{5} = 116.2$$

Al final se tienen 3 estimadores:

- $\hat{N}_1 = \max = 97$
- $\hat{N}_2 = \max + \min - 1 = 112$
- $\hat{N}_3 = \frac{(n+1)\max - 1}{n} - 116.2$

?‘Cuál de los tres estimadores es el mejor?

?‘Qué criterios debemos considerar para elegir a los ‘‘buenos’’ estimadores?

Un primer criterio para seleccionar a un estimador es que los valores cercanos al valor del parámetro sean más probables que los valores alejados a él. Como se puede ver, los estimadores $\hat{N}_1$ y $\hat{N}_3$ están en función únicamente del valor máximo en la muestra. Entonces, primero se va a encontrar la función de probabilidad de estos dos estimadores.

Función de Probabilidad de $\hat{N}_1$ y $\hat{N}_3$. Se va a encontrar la función de densidad del máximo valor en la muestra. Si se tienen N elementos numerados del 1 al N y se elige al azar sin reemplazo n de ellos, el máximo valor obtenido en la muestra puede ser cualquier número entre n y N. Esto significa que

$$\max\{X_1,\ X_2,\ \ldots,\ X_n\} = n,\ n+1,\ n+2,\ \ldots,\ N$$

Ahora vamos a calcular la probabilidad que max = x para $x = n,\ n+1,\ n+2,\ \ldots,\ N$

Si se tiene que max = x, entonces en la muestra hay $n-1$ observaciones menores que x, esto es, $n-1$ elementos de la muestra es un subconjunto de los $x-1$ datos más pequeños. El número de posibles maneras en que max = x es igual al número de subconjuntos de $n-1$ elementos de un conjunto de $x-1$ elementos.

El número de muestras tales que su máximo valor es igual a x es $\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}$

Por otro lado, el total de posibles muestras de tamaño n del total de tanques, está dado por:

$$\text{Número de muestras con } n \text{ elementos es } \begin{pmatrix} N \\ n \end{pmatrix}$$

De aquí, al aplicar la definición clásica de la probabilidad de un evento en un espacio equiprobable se tiene que:

$$P(\max = x) = \frac{\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

con $x = n,\ n+1,\ n+2,\ \ldots,\ N$

Usando esta probabilidad se puede calcular las probabilidades de los estimadores $\hat{N}_1$ y $\hat{N}_3$. Enseguida se presentan algunos ejemplos en los que se calcula esta probabilidad para casos particulares.

$$P(\hat{N}_1 = x) = \frac{\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \quad y \quad P\left(\hat{N}_3 = \frac{(n+1)x-1}{n}\right) = \frac{\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

Ejemplo 7
Considere que
$N = 20$ y
$n = 5$

El rango de $\hat{N}_1$ es $\hat{N}_1 = 5, 6, 7, \ldots, 20$ y el rango de $\hat{N}_3$ es $\hat{N}_3 = 5.8, 7, 8.2, 9.4, \ldots, 22.6, 23.8$. En la tabla y gráfica siguientes se muestran las probabilidades asociadas a $\hat{N}_1$ y $\hat{N}_3$.

max	$\hat{N}_1$	$\hat{N}_3$	Probabilidad
5	5	5.8	6.45E-05
6	6	7	0.000322
7	7	8.2	0.000967
8	8	9.4	0.002257
9	9	10.6	0.004515
10	10	11.8	0.008127
11	11	13	0.013545
12	12	14.2	0.021285
13	13	15.4	0.031927
14	14	16.6	0.046117
15	15	17.8	0.064564
16	16	19	0.088042
17	17	20.2	0.117389
18	18	21.4	0.153509
19	19	22.6	0.197368
20	20	23.8	0.250000

□

Ejemplo 8

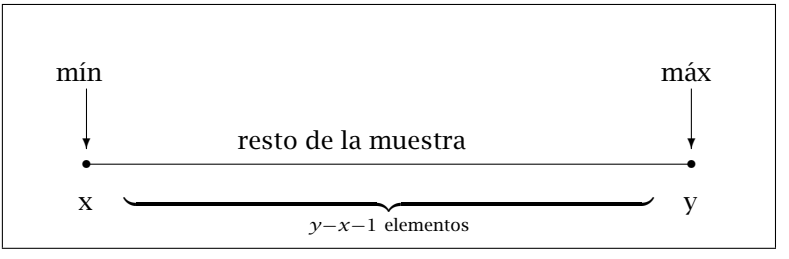
Considere que $N = 200$ y $n = 20$.

El rango de $\hat{N}_1$ es $\hat{N}_1 = 20, 21, 22, \ldots, 200$ y el rango de $\hat{N}_3$ es $\hat{N}_3 = 20.95, 22, 23.05, \ldots, 208.9, 209.95$.

En la gráfica siguientes se muestran las probabilidades del estimador $\hat{N}_1$.

□

Función de Probabilidad de $\hat{N}_2$. Para encontrar las probabilidades asociadas a $\hat{N}_2$ primero se encuentra la función de probabilidades conjunta del máximo y del mínimo valor en una muestra de tamaño n. Sean x y y tales que $x = \min\{X_1, X_2, \ldots, X_n\}$ $y = \max\{X_1, X_2, \ldots, X_n\}$, esto significa que $n-2$ elementos en la muestra están entre $x+1$ y $y-1$,



el número de muestras que satisfacen esta condición son las combinaciones de $n-2$ en $y-x-1$.

El número de muestras de tamaño n donde el máximo valor es y y el mínimo valor es x es igual a $\binom{y-x-1}{n-2}$

De esta manera se tiene que

$$P(\min = x,\ \max = y) = \frac{\binom{y-x-1}{n-2}}{\binom{N}{n}}$$

Usando la función de probabilidad conjunta, ya se puede encontrar la probabilidad correspondiente a $\hat{N}_2$.

$$P(\hat{N}_2 = m) = P(\max + \min - 1 = m) = \sum_{i=1}^{m+1-n} P(\min = i,\ \max = m-i+1)$$

$$\frac{\sum_{i=1}^{[(m-n+2)/2]} \binom{m-2i}{n-2}}{\binom{N}{n}}$$

Ejemplo 9

Considere que

$N = 20$ y

$n = 5$

Se ha calculado las probabilidades correspondientes a $\hat{N}_2$ y estas probabilidades se encuentran en una tabla. La tabla con las probabilidades y la gráfica de las mismas se presenta enseguida.

$\hat{N}_2$	Probabilidad
9	0.002966976
10	0.005159959
11	0.008384933
12	0.012899897
13	0.019027348
14	0.027089783
15	0.0374742
16	0.050567595
17	0.066821465
18	0.086687307
19	0.110681115
20	0.138318885
21	0.110681115
22	0.086687307
23	0.066821465
24	0.050567595
25	0.0374742
26	0.027089783
27	0.019027348
28	0.012899897
29	0.008384933
30	0.005159959
31	0.002966976

□

Observe que la función de probabilidad es simétrica y tiene su máximo valor cuando $\hat{N}_2 = 20$.

Ejemplo 10

Considere que

$N = 200$ y

$n = 7$

La gráfica muestra la función de probabilidad de este ejemplo.

□

La gráfica indica que el valor más probable es nuevamente $\hat{N}_2 = 20$, y la función de probabilidad también es simétrica, pero presenta menor variación que en el caso anterior.

Propiedades deseables en un estimador

Un estimador del parámetro θ es una función de los datos muestrales que ayuda a determinar el valor de θ.

Para cada parámetro pueden existir uno o más estimadores. En general, se requiere tener un estimador que posea "buenas" propiedades. Entre las propiedades deseables de un estimador se enumeran el insesgamiento, la eficiencia, la consistencia, la suficiencia y la robustez.

Insesgamiento. La idea de insesgamiento está relacionada con el concepto de exactitud. Una manera simple de explicar el concepto es la siguiente: un estimador es una función de los datos muestrales, diferentes muestras darán valores diferentes de $\hat{\theta}$. Suponga que pudiera tener los valores de $\hat{\theta}$ de todas las posibles muestras, (sean estos: $\hat{\theta}_1$, $\hat{\theta}_2$, ..., $\hat{\theta}_M$) si el promedio de todos estos valores es igual al parámetro,

$$\frac{\sum_{i=1}^{M} \hat{\theta}_i}{M} = \theta$$

entonces $\hat{\theta}$ es un estimador insesgado de θ. Si $\hat{\theta}$ es un estimador insesgado de θ, entonces θ está en el centro de los posibles valores de $\hat{\theta}$.

La definición formal de este concepto es:

Un estimador $\hat{\theta}$ del parámetro θ se dice que es insesgado si $E(\hat{\theta}) = \theta$.

EJEMPLO 2.1. Probar que el estimador de N por el método de captura y recaptura modificado dado por $\hat{N} = \frac{(n-m)(M+1)}{m} + M$ es un estimador insesgado.

Solución

En el método de captura recaptura modificado la variable aleatoria es n, y el valor esperado del estimador es igual a

$$E(\hat{N}) = E\left(\frac{(n-m)(M+1)}{m} + M\right) = \frac{(M+1)}{m}E(n-m) + M$$

Para encontrar el valor esperado de $\hat{N}$ se encuentra primero el valor esperado $E(n-m)$, y para encontrar este valor esperado se utiliza la segunda propiedad de las funciones de distribución de una variable aleatoria que en este caso es:

$$\sum_{n=m}^{N-M+m} \frac{m\binom{M}{m}\binom{N-M}{n-m}}{n\binom{N}{n}} = 1$$

para todo $m \leq M \leq N$.

Ahora, se observa que

$$E(n-m) = \sum_{n=m}^{N-M+m} \frac{(n-m)m \begin{pmatrix} M \\ m \end{pmatrix} \begin{pmatrix} N-M \\ n-m \end{pmatrix}}{n \begin{pmatrix} N \\ n \end{pmatrix}}$$

Y en esta suma se sustituyen las identidades:

$$\begin{pmatrix} M \\ m \end{pmatrix} = \frac{m+1}{M+1} \begin{pmatrix} M+1 \\ m+1 \end{pmatrix} = \frac{m^*}{M+1} \begin{pmatrix} M^* \\ m^* \end{pmatrix}$$

$$(n-m) \begin{pmatrix} N-M \\ n-m \end{pmatrix} = (N-M) \begin{pmatrix} N-M-1 \\ n-m-1 \end{pmatrix} = (N-M) \begin{pmatrix} N-M^* \\ n-m^* \end{pmatrix}$$

donde $M^* = M+1$ y $m^* = m+1$. Entonces,

$$E(n-m) = \frac{m(N-M)}{M+1} \underbrace{\sum_{n=m}^{N-M^*+m^*} \frac{m^* \begin{pmatrix} M^* \\ m^* \end{pmatrix} \begin{pmatrix} N-M^* \\ n-m^* \end{pmatrix}}{n \begin{pmatrix} N \\ n \end{pmatrix}}}_{=1} = \frac{m(N-M)}{M+1}$$

$$E(\hat{N}) = \frac{(M+1)}{m} E(n-m) + M = \frac{(M+1)}{m} \frac{m(N-M)}{M+1} + M = M$$

Eficiencia. El concepto de eficiencia está relacionado con el concepto de precisión del estimador. Primero se verá la definición de eficiencia relativa.

Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son ambos estimadores insesgados de θ, entonces si se satisface la desigualdad

$$V(\hat{\theta}_1) \leq V(\hat{\theta}_2)$$

diremos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$. Un estimador es más eficiente (más preciso), si su varianza es menor.

Un estimador θ es eficiente si $V(\hat{\theta} \leq V(\hat{\theta}^*)$ donde $\hat{\theta}^*$ es cualquier otro estimador.

Cuando se tiene un estimador insesgado del parámetro θ es útil saber si hay otro estimador insesgado más eficiente, en este sentido el teorema de Cramér-Rao proporciona una cota mínima para las varianzas de los estimadores insesgados. Aquí se presenta este teorema sin demostración.

TEOREMA 2.2 (Cramér Rao). *Si $X_1,\ X_2,\ X_3,\ \ldots,\ X_n$ es una muestra aleatoria de una función de densidad tal que $f(x;\ \theta) > 0$ para x en una región que no depende del parámetro entonces para todo estimador insesgado de θ se satisface la desigualdad*

$$V(\hat{\theta}) \geq \frac{1}{nE(\frac{\partial}{\partial\theta}\log(f(x;\ \theta)))^2}$$

Si se tiene un estimador que alcanza la cota de Cramér Rao, estaremos seguros que no hay otro estimador insesgado de θ que tenga menor varianza.

TEOREMA 2.3. *Un estimador insesgado $\hat{\theta}$ del parámetro θ es eficiente si alcanza la cota de Cramer-Rao.*

Consistencia. Si no es posible emplear estimadores de mínima varianza, el requisito mínimo deseable para un estimador, es que a medida que el tamaño de la muestra crezca, el valor del estimador tienda a estar cerca del valor del parámetro, propiedad que se denomina consistencia. Existen diversas definiciones de consistencia, más o menos restrictivas, pero la más utilizada es la siguiente.

Un estimador $\hat{\theta}$ del parámetro θ se dice que es consistente si y sólo si, cuando n crece, $\hat{\theta}$ tiende en probabilidad a θ. Esto significa que si para toda $\varepsilon > 0$ se tiene que si

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

entonces $\hat{\theta}$ es consistente.

La idea de la consistencia es que conforme el tamaño de la muestra aumenta, el valor del estimador se aproxima a la del parámetro con probabilidad igual a 1. Una condición necesaria para determinar que un estimador es consistente se presenta en el siguiente teorema

TEOREMA 2.4. *Si la varianza de un estimador insesgado es tal que*

$$\lim_{n\to\infty} V(\hat{\theta}) = 0$$

entonces $\hat{\theta}$ es un estimador consistente de θ.

Suficiencia. E1 concepto de estadística suficiente fue introducido por Fisher en 1922, y ha sido objeto de numerosas e importantes investigaciones. Como originalmente indicó Fisher, un estadístico (o función de la muestra de valores observados) es suficiente para el objetivo de la inferencia estadística si contiene, en un cierto sentido, toda la ‘‘información’’ sobre la distribución generadora (función de distribución de acuerdo con la cual ha sido generada la muestra de valores observados). ¿En qué sentido hemos de usar la palabra

"información" aquí? Supondremos que el modelo estadístico de las variables aleatorias observables tiene una cierta función de distribución conjunta que pertenece a una familia especificada $F(x;\ \theta)$ de funciones de distribución. Sin embargo, la verdadera función de distribución generadora es desconocida pues se desconoce el valor de θ.

Sea $X_1,\ X_2,\ \ldots,\ X_n$ una muestra aleatoria de la función de densidad $f(x;\ \theta)$; se llama estadística a cualquier función de los datos muestrales, esto es, cualquier función de la forma $T = h(X_1,\ X_2,\ \ldots,\ X_n)$ es una estadística.

La única información que se tiene para tomar una decisión sobre θ es el resultado del experimento aleatorio, es decir, la muestra aleatoria $X_1,\ X_2,\ \ldots,\ X_n$. Sin embargo, los datos muestrales son un complicado conjunto de números y el investigador se ve en la necesidad de introducir una simplificación deseable. Para esta simplificación, siempre que sea posible, se elegirá, un estadístico que pierda la menor información relativa al parámetro contenida en la muestra. Este es el deseo que incita a la definición de estadístico suficiente. Supóngase, pues, que T es una estadística (es decir, una función medible de la variable aleatoria X) y sea Z otra estadística; si consideramos la probabilidad condicionada de Z, dado T, en general esta probabilidad dependerá de θ; si ocurre que la función condicionada no depende del parámetro, querrá, decir que la estadística Z en presencia de T no proporciona ninguna información adicional acerca de θ. Si esto ocurre para cualquier otra estadística se concluye que toda la información que existía en la muestra $X_1,\ X_2,\ \ldots,\ X_n$ nos la ha proporcionado la estadística T, y en este caso se llama estadística suficiente.

Una estadística T se dice que es suficiente para el parámetro θ si la función de densidad condicional de U, cualquier otra estadística dada T no depende del parámetro. Esto es, si para toda U se tiene que $f_{U|T}(u|t)$ no depende de θ entonces T es estadística suficiente para θ.

En otras palabras, un estimador es suficiente para θ cuando ya tiene toda la información sobre el parámetro que está contenida en la muestra.

Es difícil probar con todas las posibles estadísticas si $f_{U|T}(u|t)$ no depende de θ, y debido a que todas las estadísticas dependen de la muestra, se puede tomar en lugar de todas las estadísticas la propia muestra $X_1,\ X_2,\ \ldots,\ X_n$. Entonces, un método para verificar si una estadística T es suficiente, es determinar la distribución condicionada de $X_1,\ X_2,\ \ldots,\ X_n$, dado T. Sin embargo, este método es también a menudo laborioso y de gran dificultad. Fisher y después Neyman proporcionaron un criterio simple, con el que podemos generalmente determinar si una familia $F(x;\ \theta)$ de funciones de distribución admite un estadístico suficiente no trivial y cuál es la forma de este estadístico.

Este criterio es proporcionado por el célebre teorema de factorización de Neyman-Fisher.

TEOREMA 2.5 (factorización de Neyman-Fisher). . *Sea T una estadística de la muestra aleatoria de la función de densidad $f(x;\ \theta)$, T es estadística suficiente del parámetro si y sólo si, la función de densidad conjunta se puede escribir como el producto de dos funciones, una que depende de la estadística suficiente y el parámetro y otra que depende únicamente de la muestra, esto es:*

$$f(x_1,\ x_2,\ \ldots,\ x_n;\ \theta) = f(x_1;\ \theta)f(x_2;\ \theta)\ldots f(x_n;\ \theta) = g(T;\ \theta)h(x_1,\ x_2,\ \ldots,\ x_n)$$

Robustez. Al estudiar un proceso aleatorio es posible que la función de distribución generadora de la muestra sea $F(x;\ \theta)$ (por ejemplo una exponencial) y se está suponiendo que la distribución correcta es $G(x;\ \theta)$, (por ejemplo una Weibull) y la estimación de θ no es afectada fuertemente por considerar que la distribución es $G(x;\ \theta)$ en lugar de $F(x;\ \theta)$, se dirá que el estimador es robusto.

Se dice que $\hat{\theta}$ es un estimador robusto del parámetro θ si los supuestos de partida en los que se basa la estimación, atribuida a la selección de la función de distribución que, en realidad, no es la correcta, no alteran de manera significativa los resultados que éste proporciona.

Método de Máxima Verosimilitud para generar Estimadores

El método de máxima verosimilitud es el método de estimación más ampliamente utilizado y se basa en escoger como estimador del parámetro al valor que maximiza la función de probabilidad dados los valores de la muestra.

Dada una muestra aleatoria $X_l,\ X_2,\ \ldots,\ X_n$, con función de densidad $f(x;\ \theta)$, se conoce como función de verosimilitud a la función de densidad conjunta de la muestra, en esta función el parámetro es la variable y la muestra se considera fija, esto es

$$L(\theta; x_1,\ x_2,\ x_3,\ \ldots,\ x_n) = f(x_1;\ \theta)f(x_2;\ \theta)f(x_3;\ \theta)\ldots f(x_n;\ \theta)\ \text{ con } \theta \in \Theta$$

Ya que las observaciones de una muestra aleatoria son independientes, la función de verosimilitud es el producto de las funciones de densidad marginales evaluada en los datos muestrales. Debido a que en la función de verosimilitud la variable es el parámetro y los valores $x_1,\ x_2,\ x_3,\ \ldots,\ x_n$ se consideran conocidos y en consecuencia, fijos, por simplicidad se utiliza la notación,

$$L(\theta) = L(\theta; x_1,\ x_2,\ x_3,\ \ldots,\ x_n) \quad \text{con } \theta \in \Theta$$

Dada una muestra aleatoria $X_1,\ X_2,\ \ldots,\ X_n$ con función de densidad $f(x;\ \theta)$, se conoce como estimador de máxima verosimilitud del parámetro θ al número donde la función de verosimilitud alcanza el valor máximo.

Entonces el objetivo del método de máxima verosimilitud es encontrar el valor dentro del posible rango de valores del parámetro que optimiza la función de verosimilitud, dados los datos de la muestra.

EJEMPLO 2.6. Encuentre el estimador de máxima verosimilitud del número de tanques que tiene el enemigo.

Para resolver este ejercicio primero se encuentra la función de verosimilitud de N dada la muestra $X_1,\ X_2,\ \ldots,\ X_n$.

Como los tanque en la muestra se obtienen sin reemplazo, una vez que se observa un número de un tanque, este ya no se vuelve a considerar para entrar en la muestra, además se supone que la selección es aleatoria y que cada tanque tiene la misma probabilidad de ser observado; en estas condiciones se tiene que

- La probabilidad que X_1 esté en la muestra es $1/N$,
- La probabilidad que X_2 esté en la muestra, dado que ya se seleccionó a X_1 es igual a $1/(N-1)$,
- La probabilidad que X_3 esté en la muestra, dado que ya se seleccionó a X_1 y a X_2 es igual a $1/(N-2)$, así se sigue hasta llegar a
- La probabilidad que X_n esté en la muestra, dado que ya se seleccionó a $X_1,\ X_2,\ \ldots,\ X_{n-1}$ es igual a $1/(N-n+1)$.

Entonces, la función de verosimilitud es igual a

$$L(N) = L(N;\ X_1,\ X_2,\ \ldots,\ X_n) = \frac{1}{N} \times \frac{1}{N-1} \times \frac{1}{N-2} \times \ldots \times \frac{1}{N-n+1}$$

Como se puede ver, la función de verosimilitud $L(N)$ no es continua, por lo que no se puede utilizar las técnicas del cálculo diferencial e integral para encontrar el valor de N que máximiza a $L(N)$. Pero haciendo un análisis de $L(N)$ se puede ver que es una función decreciente, y que $L(N) \to 0$ cuando $N \to \infty$. Por lo tanto $L(N)$ alcanza su valor más alto cuando N toma el valor más pequeño posible dada la muestra.

Por otro lado se sabe que $N \geq X_1$, $N \geq X_2$, ..., $N \geq X_n$, entonces $N \geq \max\{X_1,\ X_2,\ \ldots, X_n\}$, por lo que el valor máximo de $L(N)$ dada la muestra es cuando $N = \max\{X_1,\ X_2,\ \ldots, X_n\}$ y este es el estimador de máxima verosimilitud de N.

Ejercicios

1.- Probar que en el método de captura y recaptura el estimador $\hat{N}_4 = \frac{(n+1)(M+1)}{m+1} - 1$ es insesgado.

2.- Para $N = 100$, $M = 20$ y $n = 10$ encuentre los valores de $\hat{N}_4$ y sus probabilidades y gráfique estas probabilidades.

3.- Diga ?'cuál de los tres estimadores de N, el número de tanques es el más eficiente.

4.- Diga ?'cuál de los tres estimadores de N, el número de tanques es suficiente

5.- Diga ?'cuál de los tres estimadores de N, el número de tanques es consistente.

Capítulo 3

Prueba de Hipótesis

Las pruebas de hipótesis es una parte de la inferencia estadística que consiste, básicamente, en decidir cual de dos posibles conjeturas sobre la población es verdadera. La decisión se hace con base a la información contenida en una muestra aleatoria.

Las conjeturas relativas a la población objeto de las pruebas de hipótesis pueden estar relacionadas ó bien con la forma de la distribución de una variable aleatoria (por ejemplo si la distribución generadora de los datos es una Weibull), ó con los valores de uno o varios parámetros de la misma (por ejemplo si el número de peces es $N = 200$).

De esta manera, las pruebas de hipótesis presentan dos enfoques:

- Pruebas de hipótesis sobre parámetros, consiste en determinar si el parámetro de una distribución toma o no un determinado valor, y
- Pruebas de Bondad de Ajuste, para definir si un conjunto de datos se ha generado de una determinada distribución.

En ambos casos, se debe tomar una decisión con base en los datos de una muestra aleatoria. Con el siguiente problema se pretende explicar los conceptos de la prueba de Hipótesis sobre un parámetro.

Problema 3.
Una pelea de Campeonato de Boxeo

En una pelea de box de campeonato hay un boxeador que tiene la corona de campeón mundial de boxeo, y hay otro boxeador que está retando al campeón la corona. El retador quiere demostrar que ahora el mejor boxeador del mundo es él.

□

En este contexto, al realizarse la pelea de box se tienen dos conjeturas:

- El campeón es el mejor boxeador.
- El retador es el mejor boxeador.

Estas dos conjeturas no tienen el mismo peso, pues se da por hecho que el mejor boxeador es el campeón mundial, por lo tanto, el retador debe probar contundentemente que el es mejor que el actual campeon. Esto significa que ante un empate, la decisión de los jueces es que el campeón sigue siendo el campeón. Esto significa que la conjetura de que el campeón es el mejor boxeador tiene ventaja sobre la conjetura el retador es el mejor boxeador, pues aún cuando no se realizara la pelea, la decisión sería reconocer que el mejor es el campeón.

La conjetura que tiene ventaja, porque es verdadera, según la historia del proceso o lo que establece el modelo utilizado, se le denota como hipótesis nula. La conjetura que se contrapone a la hipótesis nula y es la que se pone a prueba es la hipótesis alternativa. En este ejemplo, se tiene que

- La hipótesis nula es: ''El campeón es el mejor boxeador''.
- La hipótesis alternativa es: ''El retador es el mejor boxeador''.

La pelea se efectúa para tomar la decisión de cual de las dos conjeturas es la verdadera se pacta la pelea. Cada round de la pelea forma parte de la muestra, en base a lo observado en la muestra se tomara la decisión.

Los criterios para decidir, entre otros, son:

- La diferencia de golpes a favor del retador.
- El número de caidas.
- la fuerza de los golpes.

Los jueces de la pelea, tienen que tomar una decisión con base en lo que observen en cada uno de los rounds. Sin embargo, es claro que al desarrollarse la pelea interviene el azar, entonces la decisión no necesariamente es acorde con la realidad, pues eventos fortuitos están presentes durante la pelea.

La siguiente tabla nos muestra las posibles relaciones entre la realidad que desconocemos y la decisión que toman los jueces de la pelea.

	La realidad es	
La decisión declara que	**El campeón es mejor**	**El retador es mejor**
El campeón es mejor	Acierto	Error tipo II
El retador es mejor	Error tipo I	Acierto

Como se puede ver, los jueces pueden decidir que el campeón es mejor o pueden decidir que el retador es mejor y por lo tanto, que hay nuevo campeón, pero la realidad no se conoce. Si la decisión es que el campeón es mejor, y en realidad el campeón es mejor, entonces se habrá tomado una decisión adecuada, lo mismo ocurre si se decide que el retador es mejor y realmente el retador es mejor. Pero, si el campeón

es mejor y se decide que el retador es mejor, se habra cometido un error (error tipo I), lo mismo ocurre si se decide que el campeón es mejor cuando en realidad el retador es mejor, en este caso se comete el error tipo II.

Lo ideal sería tener un criterio de decisión con el cual la probabilidad de cometer cualquiera de los errores, tipo I o tipo II, fuera pequeña, sin embargo esto no es posible pues si relajamos las condiciones para aceptar la hipótesis nula, se endurecen las restricciones para rechazarla, y viceversa.

Elementos de la prueba de Hipótesis. Una hipótesis estadística es una afirmación o conjetura acerca de la función de distribución $F(x,\ \theta)$ generadora de la muestra aleatoria. Si la hipótesis estadística identifica por completo la distribución, recibe el nombre de "hipótesis simple" (por ejemplo que el número de tanques es $N = 100$) y si no la especifica recibe el nombre de "hipótesis compuesta" (por ejemplo que $N > 100$).

De esta manera, se tiene que

- Una hipótesis simple es de la forma: $\sigma^2 = 100$, $N = 16$, $p = 0.5$ ó "la función generadora de la muestra es una normal estándar"
- Una hipótesis compuestas es de la forma: $N > 100$, $\sigma^2 > 16$, $p < 0.5$, ó "la función generadora de la muestra es simétrica respecto al cero".

La prueba de hipótesis esencialmente consiste en decidir entre dos conjeturas opuestas cuál es la verdadera. Una de las conjeturas se supone verdadera ya sea porque la historia o la experiencia así lo ha establecido o porque así lo indica el modelo que genera los datos. La otra conjetura es la que los datos parecen respaldar. La primera conjetura se denota como hipótesis nula y la segunda conjetura se denota como hipótesis alternativa. Generalmente, la hipótesis alternativa es la hipótesis que defiende el investigador

Se llama hipótesis alternativa a la conjetura que se pone a prueba. Es la hipótesis de investigación. La hipótesis alternativa se denota como H_a ó H_1

Se llama hipótesis nula a la conjetura que se considera cierta, ya sea porque el modelo así lo indica, porque la historia lo ha probado o porque es lo aceptado. La hipótesis nula se identifica con el símbolo H_o

Se llama estadística de prueba a la función de los datos muestrales que se utiliza para tomar la decisión.

Se llama región crítica o región de rechazo para H_o al conjunto de valores de la estadística de prueba que hace que se rechace la hipótesis nula.

Dado que la decisión se hace con base en una muestra aleatoria es posible que la decisión que se tome sea equivocada, tanto si se rechaza la hipótesis nula, como si no se rechaza. La siguiente tabla muestra la relación entre la veracidad de las hipótesis de prueba y la decisión del investigador.

		La realidad	
		H_o **es verdadera**	H_o **es falsa**
La decisión es	**No rechazar** H_o	Acierto	Error tipo II
	Rechazar H_o	Error tipo I	Acierto

Como se puede ver en la tabla, hay dos tipos error.

El error tipo I se comete al rechazar la hipótesis nula cuando es verdadera. El error tipo II se comete al no rechazar la hipótesis nula cuando es falsa.

El nivel de significancia de la prueba de hipótesis es la probabilidad de cometer el error tipo I, y se denota con la letra α.

$$\text{P(error tipo I)} = \alpha = \text{nivel de significancia de la prueba}$$

La potencia de la prueba es la probabilidad de acertar cuando se rechaza la hipótesis nula, esto es, es la probabilidad de rechazar la hipótesis nula cuando es falsa. En particular, si el parámetro toma los valores de la hipótesis alternativa se tiene que:

$$1 - P(\text{error tipo II}) = \text{potencia de la prueba.}$$

Es deseable que la región de rechazo sea tal que, la probabilidad de cometer los errores tipo I y tipo II sean pequeñas; sin embargo, no es posible disminuir la probabilidad de ambos errores simultáneamente, pues conforme disminuye uno de ellos, aumenta el otro.

Problema 4:
Aplicación de un examen

Para ingresar a cualquier universidad generalmente se aplica un examen de opción múltiple, en la mayoría de los casos, este examen consiste de 120 preguntas con 5 diferentes opciones, señaladas como a, b, c, d, e. y una solamente de estas respuestas es la correcta. Para cada aspirante a ingresar a la universidad al presentar el examen, se formula una prueba de hipótesis, donde las conjeturas son:

(1) El aspirante tiene los conocimientos suficientes para ingresar a la universidad.
(2) El aspirante no tiene los conocimientos suficientes para ingresar a la universidad.

El aspirante debe probar que tiene los conocimientos suficientes, esta es la hipótesis alternativa. La hipótesis nula es que el aspirante no tiene los conocimientos suficientes para ingresar a la Universidad

- Hipótesis nula: *Ho*: El aspirante tiene los conocimientos suficientes para ingresar a la universidad.
- Hipótesis alternativa: *Ha*: El aspirante no tiene los conocimientos suficientes para ingresar a la universidad.

La calificación del examen es la estadística de prueba y se utiliza para decidir si el aspirante es aceptado o no.

Se rechaza la hipótesis nula cuando la calificación del examen es mayor o igual a 60.

El error tipo I, (rechazar Ho cuando es verdadera) es que el estudiante pase el examen sin que tenga los conocimientos suficientes para ingresar a la Universidad.

El error tipo II (aceptar Ho cuando es falsa) es que el estudiante repruebe el examen cuando si tiene los conocimientos suficientes para ingresar a la Universidad.

La hoja de respuesta del examen es de la forma

1. (a) (b) (c) (d) (e)	50. (a) (b) (c) (d) (e)	90. (a) (b) (c) (d) (e)
2. (a) (b) (c) (d) (e)	51. (a) (b) (c) (d) (e)	91. (a) (b) (c) (d) (e)
3. (a) (b) (c) (d) (e)	52. (a) (b) (c) (d) (e)	92. (a) (b) (c) (d) (e)
4. (a) (b) (c) (d) (e)	53. (a) (b) (c) (d) (e)	93. (a) (b) (c) (d) (e)
⋮	⋮	⋮

La probabilidad de elegir una respuesta correcta es p, conforme más preparado está el estudiante, mayor será la probabilidad de acertar en cada pregunta. La calificación del examen es una variable aleatoria cuya función de probabilidad es binomial con parámetros $n = 120$ y p. Si el aspirante no tiene ningún conocimiento y solo está contestando al azar, se tendrá que $p = 1/5 = 0.20$. Si el estudiante conoce el tema se tendrá que $p > 0.20$

Entonces las hipótesis de prueba se pueden escribir así,

$$Ho: \ p = 0.20 \quad \text{contra} \quad Ha: \ p > 0.20$$

$$\text{P(error tipo I) = P(que el estudiante sea admitido cuando no está preparado)}$$

$$= P(Calificación \geq 60 \mid p = 1/5) \simeq 0$$

Esta probabilidad se calcula considerando la función de probabilidad binomial.

Como se ve, es casí imposible que un estudiante que no sabe nada, pueda obtener más de 60 aciertos.

Si el aspirante tiene algún conocimiento, la probabilidad de acertar p es mayor a 0.20, pero cual es la potencia de la prueba para los diferentes valores de $p > 0.20$

La potencia de la prueba es

$$\begin{aligned} Potencia &= 1 - P(errorII) \\ &= 1 - P(Calificación < 60 | p > 0.20) \end{aligned}$$

En la siguiente figura se muestra la potencia de la prueba, en función de la probabilidad p.

□

Como se puede ver, conforme mayor sea la probabilidad de acertar, aumenta la potencia de la prueba.

Si se quiere disminuir la probabilidad de que alguien que no sabe pase el examen, se puede auemtar el puntaje de acreditación, esto significa que para pasar el examen se puede pedir que la calificación sea mayor o igual a 80. Es más difícil que alguien que no está preparado sea admitido, sin embargo aumenta la probabilidad que sea rechaado alguien que si está preparado. Esto significa que al disminuir la probabilidad de uno de los errores, aumenta la probabilidad del otro error.

En resumen: Una prueba o contraste de una hipótesis estadística es una regla o procedimiento que conduce a la decisión de aceptar o rechazar cierta hipótesis, identificada como hipótesis nula, con base en los resultados de una muestra. Los procedimientos de prueba de hipótesis dependen del empleo de la información contenida en una muestra aleatoria de la población de interés. Si esta información es consistente con la hipótesis nula se concluye que esta es verdadera; sin embargo, si esta información es inconsistente con la hipótesis nula se concluye que esta es falsa.

Los pasos a seguir en una prueba de hipótesis es:

(1) Formular las hipótesis.
(2) Tomar una muestra aleatoria de la variable de interés $X_1, X_2, \ldots, X_n$.
(3) Generar o calcular un "Estadístico de prueba", que servirá para definir la acción de aceptar o rechazar la hipótesis nula.
(4) Definir el criterio de aceptación o de rechazo. Es decir, el procedimiento de prueba parte los posibles valores del estadístico de prueba en dos subconjuntos o regiones: Una "región de aceptación de H_0" y una "región de rechazo de H_0".

(5) Tomar la decisión de aceptar o rechazar H_0 dependiendo de si el estadístico de prueba queda en la región de aceptación o en la región de rechazo.

Es importante comprender que la aceptación de una hipótesis nula simplemente implica que los datos obtenidos no dan suficiente evidencia para rechazarla. Por otro lado, el rechazo de una hipótesis implica que la evidencia muestral pone en duda la hipótesis planteada.

El lema de Neuman Pearson proporciona una forma de obtener "la mejor" región de rechazo, esto es, la región de rechazo que tiene menor probabilidad de error II, fijando la probabilidad del error I.

TEOREMA 3.1 (Lema de Neuman Pearson). *Dadas las hipótesis simples*

$$H_0 : \theta = \theta_0 \quad \text{contra} \quad H_1 : \theta = \theta_1$$

la región dada por

$$C = \{X_1, X_2, \ldots, X_n \mid \frac{L(\theta_0}{L(\theta_1)} < \lambda\}$$

es la mejor región crítica, en el sentido que

- $P(errorI) = P(C \mid \theta = \theta_0) = \alpha$ *y*
- $P(errorII) = P(C \mid \theta = \theta_1) = \beta$ *con* β *mínima.*

Ejercicio

1.- Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria Bernoulli con parámetro p, $(f(x) = p^x(1-p)^{1-x}$, para $x = 0, 1)$. Encuentre la mejor región crítica para las hipótesis

$$H_0 : p = 0.4 \quad \text{contra} \quad H_a : p = 0.8$$

2.- Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria Poisson con parámetro p, $(f(x) = \frac{\lambda^x e^{-\lambda}}{x!})$. Encuentre la mejor región crítica para las hipótesis

$$H_0 : \lambda = 5 \quad \text{contra} \quad H_a : \lambda = 2$$

Capítulo 4

Estimación de parámetros

Como ya se dijo, las técnicas de inferencia estadística caen en una de dos posibles categorías: **estimación de parámetros** y **pruebas de hipótesis estadísticas**. En este capítulo se revisarán dos problemas en los que se introduciran los elementos de la estimación de parámetros.

Un parámetro es un valor relacionado con la población objetivo y por lo regular es desconocido. Por ejemplo, el promedio del ingreso en la población, el número de elementos en la población, etc.

Un estimador del parámetro es una función de los valores muestrales, con el cual pretendemos conocer el valor del parámetro. Por ejemplo, el promedio de una muestra estima el valor de la media de la población.

La estadística nos proporciona métodos para obtener estimadores y criterios para establecer las propiedades deseables en los estimadores.

Problema 1.
¿Cuántos Peces Hay en el Lago?

El guardabosques de un parque nacional quiere saber cuántas carpas adultas habitan en el lago del parque, pues debe decidir si se permite la pesca deportiva o si se declara un estado de veda.

El guardia llega a la conclusión que no es posible contar todas las carpas directamente; por lo que debe idear la manera de obtener indirectamente una estimación aceptable de N el total de carpas adultas en el lago. El guardia considera que en el lago hay N carpas adultas. N es un número asociado a la población, y por lo tanto es el **parámetro poblacional** y es desconocido.

La distribución uniforme de los peces en el lago. Después de cavilar un rato, el guardia considera haber encontrado una posible solución; como se tienen los medios técnicos para determinar con suficiente precisión el volumen total del lago V (por ejemplo con ultrasonido) piensa que si delimita una pequeña región en el lago, de

volumen v, y cuenta los peces que hay en ese espacio, entonces puede utilizar esta información para obtener un estimador de N.

Entonces se tendrían los siguientes valores,

- N es el número de carpas adultas en el lago;
- V es el volumen del lago;
- v es el volumen en la región controlada por el guardia;
- n es el número de carpas adultas en el volumen v.

El guardia forestal considera que el número de peces en la región pequeña controlada, es proporcional al número de peces en el volumen total del lago.

Bajo este supuesto de proporcionalidad, se satisface la relación:

$$\frac{N}{V} = \frac{n}{v}$$

de donde se puede despejar el valor de N

$$N = \frac{n}{v}V$$

Una vez modelado el problema y con una fórmula para resolverlo cabe preguntarse:

¿Qué tan válido es este valor como estimador de N?. Para deducir esta fórmula, se considero que los peces se encuentran uniformemente distribuidos en todo el lago. Pues si el número de peces en el lago es proporcional al número de peces en la región controlada es porque en todo el lago están los peces distribuidos uniformemente, como se observa en la figura (a).

(a) Distribución uniforme de los peces en todo el lago	(b) Distribución de los peces en forma de cardúmenes

Sin embargo; esta supocisión no es cierta, ya que los peces se encuentran reunidos en cardúmenes y no diseminados uniformemente por todo el lago como se observa en la figura (b).

De ahí que el supuesto que el número de carpas en la región controlada, sea proporcional al número total de peces en el lago es falso y entonces la ecuación de proporcionalidad no corresponde a la realidad.

Además la presencia del guardia puede alterar la conducta de los peces, los peces pueden huir al ver al guardia, o pueden reaccionar acercándose al mismo. y puede ser difícil contarlos dentro de la región controlada de volumen v.

Esto significa que al contrastar la solución propuesta con la realidad los resultados son insatisfactorios. De manera que ni es fácil tomar la muestra, ni la suposición del modelo es acorde con la realidad.

Es decir: el modelo no es una buena aproximación con la realidad y resulta inadecuado.

El guardia necesita proponer otro modelo que sea acorde con la conducta de los peces.

Método de captura y recaptura. Este método de solución tiene dos etapas:

- **Primera Etapa**: se sacan del lago M carpas adultas, se les marca con una señal y se regresan vivas al lago; después de esto, se deja un tiempo razonable para que las carpas marcadas se integren al resto de las carpas.

  Al final de esta primera etapa, en el lago existen dos clases de carpas, las carpas marcadas y las carpas sin marcar.

- **Segunda Etapa**: se extraen al azar n carpas adultas del lago (a este subconjunto de carpas se denomina "la muestra de tamaño n", y luego se cuenta el número de carpas marcadas que contiene esta muestra. Suponga que en la muestra de n carpas se observan m carpas marcadas.

Se tiene entonces que:

- N es el número total de carpas adultas en el lago:
- M es el número total de carpas marcadas en el lago.
- n es el número total de carpas en la muestra.
- m es el número de carpas marcadas en la muestra.

Bajo el supuesto que las carpas marcadas y las no marcadas se encuentran mezcladas uniformemente en el cardumen, se puede suponer que el número de carpas marcadas en la muestra es proporcional al número de carpas marcadas en la población total, esto es, se debe satisfacer la relación:

$$\frac{N}{M} = \frac{n}{m} \tag{4.1}$$

de donde al despejar el valor de N, se obtiene

$$N = \frac{n}{m} M \tag{4.2}$$

Igual que antes, cabe preguntarse sobre que tan bueno es este número como estimador del parámetro N.

Al establecer la relación 4.1, se consideró que los peces marcados se encuentran uniformemente distribuidos con el resto de la población total.

Este supuesto a nuestro entender se cumple razonablemente, pues no es creible que los peces marcados estén todos juntos formando un grupo sin mezclarse con el resto de la población de peces; por lo que consideramos que el modelo propuesto parece ser adecuado para estimar a N.

Pero, a pesar de ser éste un modelo razonablemente adecuado, el valor obtenido con la fórmula 4.2 no corresponde al verdadero valor de N, porque en el proceso del muestreo interviene el azar, esto es, el número de peces marcados en la muestra es una variable aleatoria, y puede cambiar si se otra muestra de tamaño n en circunstancias semejantes.

Para distinguir el estimador del valor real (o parámetro), al estimador de N y se denota con la misma letra N pero con un gorrito arriba.

$$\hat{N} = \frac{n}{m} M \tag{4.3}$$

N es el parámetro poblacional y $\hat{N}$ es el estimador de N.

Como ejemplo se va a revisar el caso en que en total se tiene 20 peces y que de estos, 4 están marcados; es decir, N=20 y M=4.

Si se elijen 10 de estos peces al azar, en la muestra puede haber desde cero a cuatro peces marcados, y únicamente cuando en la muestra hay 2 peces marcados es que se satisface satisface la relación 4.1,

$$\frac{20}{4} = \frac{10}{2}$$

y por lo tanto, la estimación coincide con el del parámetro, esto es $\hat{N} = 4 \times 10/2 = 20$.

El número de peces marcados en la muestra, X, es un número aleatorio que puede ir de cero a cuatro ($X = 0,\ 1,\ 2,\ 3,\ 4$) por lo tanto los valores que puede tomar el estimador $\hat{N} = 4 \times 10/x$ son los siguientes:

- Cuando $X = 0$ se tiene que $\hat{N} = \infty$
- Cuando $X = 1$ se tiene que $\hat{N} = 40$
- Cuando $X = 2$ se tiene que $\hat{N} = 20$
- Cuando $X = 3$ se tiene que $\hat{N} = 13.33$
- Cuando $X = 4$ se tiene que $\hat{N} = 10$

El estimador de N es una variable aleatoria y podemos preguntarnos con que probabilidad puede tomar los diferentes valores posibles.

La probabilidad se puede calcular usando la definición clásica.

$$P(X = m) = \frac{Número\ total\ de\ muestras\ de\ tamaño\ n\ con\ m\ peces\ marcados}{Número\ total\ de\ muestras\ de\ tamaño\ n}$$

De esta manera nos encontramos ante el problema de tener que contar, lo cual nos lleva a recordar el cálculo combinatorio, en particular la definición de combinaciones.

Número total de muestras de tamaño n (número de elementos del espacio muestral). Observe que la muestra es un subconjunto de n elementos del conjunto total que tiene N elementos, entonces para saber cuántas muestras de tamaño n es posible obtener, se utiliza la fórmula de las combinaciones, esto es

$$\text{El total de muestras de tamaño } n = \binom{N}{n}$$

Ahora se va a obtener una fórmula para determinar cuántas de estas posibles muestras tienen m peces marcados.

Número total de muestras de tamaño n con m peces marcados (Casos favorables). Del total de posibles muestras de tamaño n, interesa determinar cuántas de ellas tienen exactamente m carpas marcadas, para ello se debe ver que las m carpas marcadas en la muestra, es un subconjunto de las M carpas marcadas en el lago, y que las $n - m$ carpas no marcadas en la muestra, es un subconjunto de las $N - M$ carpas no marcadas en el lago, de aquí se sigue que:

El total de subconjuntos de carpas marcadas de las M que hay en el lago es: $\binom{M}{m}$

El total de subconjuntos de $n - m$ carpas no marcadas de las $N - M$ que hay en el lago es: $\binom{N-M}{n-m}$

Por la regla del producto de los métodos de conteo, se sigue que el total de muestras de tamaño n con exactamente m carpas marcadas es el producto de estos dos números, esto es:

$$\text{El total de muestras de tamaño } n \text{ con } m \text{ carpas marcadas} = \binom{M}{m}\binom{N-M}{n-m}$$

entonces la probabilidad de tener una muestra con exactamente $X = m$ carpas marcadas es igual al cociente:

$$P(X = m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

Ahora se presentarán cuatro ejemplos para explorar el comportamiento de esta fórmula en función de la variable m.

Ejemplo 1

Considere que

$N = 20$

$M = 8$ y

$n = 6$

Si en la segunda etapa de captura se eligen 6 peces, El valor de m (el número de peces marcados en la muestra) puede tomar los valores de 0 a 6.

En cada caso se tendría un valor del estimador con una cierta probabilidad. En la siguiente tabla se muestra los valores de m, los valores de $\hat{N}$ y las respectivas probabilidades,

m	$\hat{N}$	Probabilidad
0	∞	0.02383901
1	48	0.16346749
2	24	0.35758514
3	16	0.31785346
4	12	0.11919505
5	9.6	0.01733746
6	8	0.00072239

En la tabla se observa que los valores del estimador que se encuentran más cerca a $N = 20$, son 16 y 24. La altura de las barras de la gráfica corresponden a la probabilidad de cada valor. Se observa que los valores 16 y 24 son los mas probables y juntos tienen una probabilidad igual a 0.675438. También se puede observar que los posibles valores del estimador tienen una gran variabilidad, pues va de 8 a ∞.

Ejemplo 2

$N = 20$,

$M = 8$ y

$n = 5$.

En este caso, el rango de X es de 0 a 5 ($X = 0,\ 1,\ 2,\ 3,\ 4$ y 5) y cuando $X = 2$ se satisface la ecuación 4.1 por lo que la estimación coincide con el verdadero valor de $N = 20$. En la siguiente tabla se presentan los valores del estimador y sus probabilidades respectivas,

m	$\hat{N}$	Probabilidad
0	∞	0.05108359
1	40	0.25541796
2	20	0.39731682
3	13.333	0.23839009

Observe que la barra asociada al número 20 es la más alta, lo que indica que este valor es el más probable.

También se puede observar que los posibles valores del estimador tienen una gran variabilidad, pues va de 8 a ∞.

Ejemplo 3
$N = 100$,
$M = 20$ y
$n = 10$.

El valor de m (el número de peces marcados en la muestra) puede tomar los valores de 0 a 10. En cada caso se tendría un valor del estimador con una cierta probabilidad. En la siguiente tabla se muestra los valores de m, los valores de $\hat{N}$ y las respectivas probabilidades,

□

m	$\hat{N}$	Probabilidad
0	∞	0.09511627
1	200	0.26793316
2	100	0.31817063
3	66.667	0.20920809
4	50	0.0841073
5	40	0.02153147
6	33.333	0.00354136
7	28.571	0.00036793
8	25	2.2996E-05
9	22.222	7.7623E-07
10	20	1.0673E-08

El mejor valor para estimar a $N = 100$ es cuando $m = 2$, porque en este caso $\hat{N} = 100$. También en este caso, es el valor más probable. La variación de los valores del estimador tiene una variación muy grande, va del 20 al infinito.

Ejemplo 4
$N = 500$,
$M = 100$ y
$n = 50$.

El rango de X es 0 a 50 ($X = 0, 1, 2, 3, \ldots, 50$) y una tabla con 51 entradas es muy grande para escribirla aquí, por lo que sólo se presenta la gráfica con las probabilidades de los valores que puede tomar el estimador.

□

También en este ejemplo $\hat{N} = 500$ es el valor más probable y el estimador tiene gran variación.

En los cuatro ejemplos revisados se observó que el valor del estimador más cercano al verdadero valor del parámetro es el más probable; sin embargo, los posibles valores de la estimación presentan una gran variación, por lo que el estimador podría estar muy lejos del parámetro que estiman.

Método de captura y recaptura modificado. El estimador $\hat{N} = nM/m$ tiene el inconveniente de que se requiere dividir entre 0, cuando $m = 0$, esto provoca que su dispersión sea muy grande.

Este inconveniente se puede eliminar si modificamos el esquema de muestreo.

Este nuevo método tiene también dos etapas, la primera es exactamente igual que la del método anterior, para tener peces marcados y no marcados en el lago.

- **Primera Etapa**: se sacan del lago M carpas adultas, se les marca con una señal y se regresan vivas al lago; después de esto, se deja un tiempo razonable para que las carpas marcadas se integren al resto de las carpas.
- **Segunda Etapa**: se extraen al azar sucesivamente una carpa tras otra y se detiene el procedimiento cuando se obtengan m carpas marcadas (m un número preestablecido).

En este caso el total de observaciones es aleatorio. Si el número de extracciones es n significa que con la extracción $n-1$ ya se tienen en la muestra $m-1$ peces marcados, y que en la extracción siguiente se obtiene el último pez marcado con el que se completa los m peces en la muestra.

□

En este punto se tiene que:

- N es el número total de carpas adultas en el lago:
- M es el número total de carpas marcadas en el lago.
- n es el número total de carpas extraidas.
- m es el número de carpas marcadas en la muestra.

El número n es aleatorio y m es fijo, pues de antemano se indica cuántos peces marcados se quieren en la muestra y no se sabe apriori el número de extracciones necesarias para seleccionar exactamente m carpas marcadas.

La proporción que se propone en este caso es

$$\frac{N-M}{M+1} = \frac{n-m}{m}$$

y al despejar el valor de N se obtiene un nuevo estimador

$$\hat{N} = \frac{(n-m)(M+1)}{m} + M$$

Los valores $M+1$ y m son constantes conocidas en esta fórmula. Cuando m divide a $M+1$ los valores del estimador son todos enteros, si esto no ocurre, los valores del estimador pueden no ser enteros. Ahora se analizara la probabilidad de este estimador.

Casos totales. La extracción de las carpas se realiza sucesivamente hasta tener m carpas marcadas y para determinar la cardinalidad del espacio muestral se consideran dos etapas: la primera es cuando se

obtienen las primeras $n-1$ extracciones, la segunda etapa es cuando se obtiene la última carpa que debe estar marcada.

Los primeros $n-1$ peces seleccionados forman un subconjunto de la población total de carpas, y el total de formas en que se pueden elegir es igual a $\binom{N}{n-1}$ y el último pez seleccionado puede ser cualquiera de los $N-n+1$ que quedan en el lago. Por la regla del producto de los métodos de conteo, se tiene que el total de formas de extraer n carpas del total de carpas, siendo la última una carpa marcada es:

$$\binom{N}{n-1}(N-n+1)$$

Casos Favorables. Del total de posibles muestras de tamaño n interesa determinar cuántas tienen exactamente $m-1$ carpas marcadas en las $n-1$ primeras extracciones.

Las $m-1$ carpas marcadas es un subconjunto de las M carpas marcadas, y las $n-m$ carpas no marcadas es un subconjunto de las $N-M$ carpas no marcadas.

El total de subconjuntos de $m-1$ carpas marcadas de las M carpas marcadas es igual a $\binom{M}{m-1}$

El total de subconjuntos de $n-m$ carpas no marcadas del total de $N-M$ carpas no marcadas es igual a $\binom{N-M}{n-m}$

La última carpa marcada puede ser cualquiera de las $M-m+1$ carpas marcadas que permanecen aún en el lago. Por la regla del producto de los métodos de conteo, el total de muestras con exactamente m carpas marcadas es el producto de estos tres terminos, esto es:

El total de muestras de tamaño n con m carpas marcadas es

$$\binom{M}{m-1}\binom{N-M}{n-m}(M-m+1)$$

De donde se tiene que la probabilidad de detenerse en la extracción n, $(X=n)$, es igual al cociente

$$P(X=n)=\frac{\binom{M}{m-1}\binom{N-M}{n-m}(M-m+1)}{\binom{N}{n-1}(N-n+1)}=\frac{m\binom{M}{m}\binom{N-M}{n-m}}{n\binom{N}{n}}$$

Para estudiar esta fórmula es conveniente verla gráficamente para algunos casos particulares.

Ejemplo 5

Considere que
$N = 207$,
$M = 8$ y
$m = 3$ (con esto se tiene que $(M+1)/m$ es un entero

En este caso los posibles valores que puede tomar la variable n es de 3 a 15 ($n = 3,\ 4,\ 5,\ \ldots,\ 15$); en la siguiente tabla se presentan los valores del estimador y sus probabilidades,

n	Estimador de N	Probabilidad
3	8	0.049123807
4	11	0.104024768
5	14	0.143034056
6	17	0.158926729
7	20	0.153250774
8	23	0.132031436
9	26	0.102691117
10	29	0.072017147
11	32	0.045010717
12	35	0.024436260
13	38	0.01100262
14	41	0.00371517
15	44	0.000722394

□

Se puede observar que el mejor valor de $\hat{N}$ es cuando $\hat{N} = 20$, este valor tiene una probabilidad alta de ocurrir. Además se puede ver que este estimador tiene menor variación que con el esquema original de muestreo de captura y recaptura.

Ejemplo 6

Considere que
$N = 100$,
$M = 20$ y
$m = 3$ (con esto se tiene que $(M+1)/m$ es un entero

En este caso los posibles valores que puede tomar la variable n es de 3 a 83 ($n = 3,\ 4,\ 5,\ \ldots,\ 83$); en la siguiente gráfica se presentan los valores del estimador y sus probabilidades.

□

Se observa que alrededor de 100 se encuentran los valores más probables.

Conclusion: Los dos esquemas de muestreo proporcionan estimadores de N cuyos valores más cercanos al valor real de N son más probables, sin embargo, el método de captura y recaptura modificado proporciona estimadores con menos dispersión.

Problema 2
¿Cuántos tanques tiene el enemigo?

Ahora se va a estudiar un problema diferente. Se considera dos ejercitos que están en guerra y los estrategas de uno de los ejercitos quieren determinar el número de tanques que tiene el otro ejercito.

□

Es claro que se puede conocer este número revisando los inventarios del otro ejercito, pero esto no es posible y por lo tanto se debe recurrir a un método indirecto.

□

Se ha visto que los tanques observados del otro ejercito traen escrito un número, lo que sugiere que todos los tanques deben estar numerados del 1 al N, siendo N el total de tanques.

□

Además, se puede considerar que es igualmente probable observar a cualquiera de los tanques del enemigo. El reto es entonces estimar el valor de N.

□

De esta manera se observa el número que traen escrito 5 tanques que fueron observados en diferentes lugares, los números observados son: 81, 47, 16, 97, 58. Estos datos son nuestra muestra aleatoria.

□

Con estos cinco números ?'qué podemos decir del valor de N?

Para tener diferentes opciones, la tarea de determinar el valor de N se la encomiendan a tres grupos diferentes que trabajaran independientemente.

Para estimar el número N, primero se ordenan los datos de la muestra.

$$16,\ 47,\ 58,\ 81,\ 97$$

Propuesta de estimación del equipo 1

El primer equipo de estrategas considera que si los tanques están numerados sucesivamente entonces N debe ser mayor o igual a 97. Y entonces proponen como estimador, precisamente a este número.

$$\hat{N}_1 = \max\{X_1,\ X_2,\ X_3,\ X_4,\ X_5\} = \max = 97$$

El primer equipo estima que el ejercito contrario tiene 97 tanques.

Propuesta de estimación del equipo 2

El segundo equipo de estrategas también considera que si los tanques están numerados sucesivamente entonces N debe ser mayor o igual a 97, pero piensa que si el tanque número uno no fue observado, bien podría ser que tampoco se hubiera observado el último tanque (el marcado con el número N), y para describir esto escriben así nuevamente la lista de datos:

$$1,\ 16,\ 47,\ 58,\ 81,\ 97,\ N$$

Este equipo considera que la separación entre max = 97 y N debe ser similar a la separación entre min = 16 y 1, esto es

$$N - \max = \min - 1 \ \Rightarrow\ N = \max + \min - 1$$

El estimador propuesto por el segundo equipo de estrategas es

$$\hat{N}_2 = \max + \min - 1 = 97 + 16 - 1 = 112$$

Propuesta de estimación del equipo 3

El tercer equipo de estrategas también considera que si los tanques están numerados sucesivamente entonces N debe ser mayor o igual a 97, y razonan que las separaciones entre dos números observados debe ser semejante, por lo que calculan el promedio de las separaciones entre dos datos consecutivos.

$$1, \quad 16, \quad 47, \quad 58, \quad 81, \quad 97, \quad N$$

$$d_1 \quad d_2 \quad d_3 \quad d_4 \quad d_5 \quad d$$

Proponen que

$$d = \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5}$$

$$= \frac{(16-1) + (47-16) + (58-47) + (81-58) + (97-81)}{5} = \frac{97-1}{5}$$

Esto lleva a la ecuación

$$N - \max = d \quad \Rightarrow \quad N = \max + d = \frac{6\max - 1}{5}$$

Y el estimador propuesto por el tercer equipo es

$$\hat{N}_3 = \frac{(n+1)\max - 1}{n} = \frac{6 \times 97 - 1}{5} = 116.2$$

Al final se tienen 3 estimadores:

- $\hat{N}_1 = \max = 97$
- $\hat{N}_2 = \max + \min - 1 = 112$
- $\hat{N}_3 = \frac{(n+1)\max - 1}{n} - 116.2$

?‘Cuál de los tres estimadores es el mejor?

?‘Qué criterios debemos considerar para elegir a los ‘‘buenos’’ estimadores?

Un primer criterio para seleccionar a un estimador es que los valores cercanos al valor del parámetro sean más probables que los valores alejados a él. Como se puede ver, los estimadores $\hat{N}_1$ y $\hat{N}_3$ están en función únicamente del valor máximo en la muestra. Entonces, primero se va a encontrar la función de probabilidad de estos dos estimadores.

Función de Probabilidad de $\hat{N}_1$ y $\hat{N}_3$. Se va a encontrar la función de densidad del máximo valor en la muestra. Si se tienen N elementos numerados del 1 al N y se elige al azar sin reemplazo n de ellos, el máximo valor obtenido en la muestra puede ser cualquier número entre n y N. Esto significa que

$$\max\{X_1,\ X_2,\ \ldots,\ X_n\} = n,\ n+1,\ n+2,\ \ldots,\ N$$

Ahora vamos a calcular la probabilidad que max = x para $x = n,\ n+1,\ n+2,\ \ldots,\ N$

Si se tiene que max = x, entonces en la muestra hay $n-1$ observaciones menores que x, esto es, $n-1$ elementos de la muestra es un subconjunto de los $x-1$ datos más pequeños. El número de posibles maneras en que max = x es igual al número de subconjuntos de $n-1$ elementos de un conjunto de $x-1$ elementos.

El número de muestras tales que su máximo valor es igual a x es $\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}$

Por otro lado, el total de posibles muestras de tamaño n del total de tanques, está dado por:

$$\text{Número de muestras con } n \text{ elementos es } \begin{pmatrix} N \\ n \end{pmatrix}$$

De aquí, al aplicar la definición clásica de la probabilidad de un evento en un espacio equiprobable se tiene que:

$$P(\max = x) = \frac{\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

con $x = n,\ n+1,\ n+2,\ \ldots,\ N$

Usando esta probabilidad se puede calcular las probabilidades de los estimadores $\hat{N}_1$ y $\hat{N}_3$. Enseguida se presentan algunos ejemplos en los que se calcula esta probabilidad para casos particulares.

$$P(\hat{N}_1 = x) = \frac{\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \quad y \quad P\left(\hat{N}_3 = \frac{(n+1)x-1}{n}\right) = \frac{\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

Ejemplo 7
Considere que
$N = 20$ y
$n = 5$

El rango de $\hat{N}_1$ es $\hat{N}_1 = 5, 6, 7, \ldots, 20$ y el rango de $\hat{N}_3$ es $\hat{N}_3 = 5.8, 7, 8.2, 9.4, \ldots, 22.6, 23.8$. En la tabla y gráfica siguientes se muestran las probabilidades asociadas a $\hat{N}_1$ y $\hat{N}_3$.

max	$\hat{N}_1$	$\hat{N}_3$	Probabilidad
5	5	5.8	6.45E-05
6	6	7	0.000322
7	7	8.2	0.000967
8	8	9.4	0.002257
9	9	10.6	0.004515
10	10	11.8	0.008127
11	11	13	0.013545
12	12	14.2	0.021285
13	13	15.4	0.031927
14	14	16.6	0.046117
15	15	17.8	0.064564
16	16	19	0.088042
17	17	20.2	0.117389
18	18	21.4	0.153509
19	19	22.6	0.197368
20	20	23.8	0.250000

□

Ejemplo 8

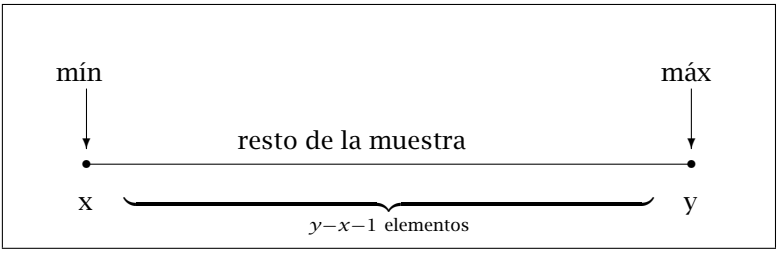
Considere que $N = 200$ y $n = 20$.

El rango de $\hat{N}_1$ es $\hat{N}_1 = 20, 21, 22, \ldots, 200$ y el rango de $\hat{N}_3$ es $\hat{N}_3 = 20.95, 22, 23.05, \ldots, 208.9, 209.95$.

En la gráfica siguientes se muestran las probabilidades del estimador $\hat{N}_1$.

□

Función de Probabilidad de $\hat{N}_2$. Para encontrar las probabilidades asociadas a $\hat{N}_2$ primero se encuentra la función de probabilidades conjunta del máximo y del mínimo valor en una muestra de tamaño n. Sean x y y tales que $x = \min\{X_1, X_2, \ldots, X_n\}$ $y = \max\{X_1, X_2, \ldots, X_n\}$, esto significa que $n-2$ elementos en la muestra están entre $x+1$ y $y-1$,



el número de muestras que satisfacen esta condición son las combinaciones de $n-2$ en $y-x-1$.

El número de muestras de tamaño n donde el máximo valor es y y el mínimo valor es x es igual a $\binom{y-x-1}{n-2}$

De esta manera se tiene que

$$P(\min = x,\ \max = y) = \frac{\binom{y-x-1}{n-2}}{\binom{N}{n}}$$

Usando la función de probabilidad conjunta, ya se puede encontrar la probabilidad correspondiente a $\hat{N}_2$.

$$P(\hat{N}_2 = m) = P(\max + \min - 1 = m) = \sum_{i=1}^{m+1-n} P(\min = i,\ \max = m-i+1)$$

$$\frac{\sum_{i=1}^{[(m-n+2)/2]} \binom{m-2i}{n-2}}{\binom{N}{n}}$$

Ejemplo 9

Considere que

$N = 20$ y

$n = 5$

Se ha calculado las probabilidades correspondientes a $\hat{N}_2$ y estas probabilidades se encuentran en una tabla. La tabla con las probabilidades y la gráfica de las mismas se presenta enseguida.

$\hat{N}_2$	Probabilidad
9	0.002966976
10	0.005159959
11	0.008384933
12	0.012899897
13	0.019027348
14	0.027089783
15	0.0374742
16	0.050567595
17	0.066821465
18	0.086687307
19	0.110681115
20	0.138318885
21	0.110681115
22	0.086687307
23	0.066821465
24	0.050567595
25	0.0374742
26	0.027089783
27	0.019027348
28	0.012899897
29	0.008384933
30	0.005159959
31	0.002966976

□

Observe que la función de probabilidad es simétrica y tiene su máximo valor cuando $\hat{N}_2 = 20$.

Ejemplo 10

Considere que

$N = 20$ y

$n = 7$

La gráfica muestra la función de probabilidad de este ejemplo.

□

La gráfica indica que el valor más probable es nuevamente $\hat{N}_2 = 20$, y la función de probabilidad también es simétrica, pero presenta menor variación que en el caso anterior.

Propiedades deseables en un estimador

Un estimador del parámetro θ es una función de los datos muestrales que ayuda a determinar el valor de θ.

Para cada parámetro pueden existir uno o más estimadores. En general, se requiere tener un estimador que posea "buenas" propiedades. Entre las propiedades deseables de un estimador se enumeran el insesgamiento, la eficiencia, la consistencia, la suficiencia y la robustez.

Insesgamiento. La idea de insesgamiento está relacionada con el concepto de exactitud. Una manera simple de explicar el concepto es la siguiente: un estimador es una función de los datos muestrales, diferentes muestras darán valores diferentes de $\hat{\theta}$. Suponga que pudiera tener los valores de $\hat{\theta}$ de todas las posibles muestras, (sean estos: $\hat{\theta}_1$, $\hat{\theta}_2$, ..., $\hat{\theta}_M$) si el promedio de todos estos valores es igual al parámetro,

$$\frac{\sum_{i=1}^{M} \hat{\theta}_i}{M} = \theta$$

entonces $\hat{\theta}$ es un estimador insesgado de θ. Si $\hat{\theta}$ es un estimador insesgado de θ, entonces θ está en el centro de los posibles valores de $\hat{\theta}$.

La definición formal de este concepto es:

Un estimador $\hat{\theta}$ del parámetro θ se dice que es insesgado si $E(\hat{\theta}) = \theta$.

EJEMPLO 4.1. Probar que el estimador de N por el método de captura y recaptura modificado dado por $\hat{N} = \frac{(n-m)(M+1)}{m} + M$ es un estimador insesgado.

Solución

En el método de captura recaptura modificado la variable aleatoria es n, y el valor esperado del estimador es igual a

$$E(\hat{N}) = E\left(\frac{(n-m)(M+1)}{m} + M\right) = \frac{(M+1)}{m}E(n-m) + M$$

Para encontrar el valor esperado de $\hat{N}$ se encuentra primero el valor esperado $E(n-m)$, y para encontrar este valor esperado se utiliza la segunda propiedad de las funciones de distribución de una variable aleatoria que en este caso es:

$$\sum_{n=m}^{N-M+m} \frac{m\binom{M}{m}\binom{N-M}{n-m}}{n\binom{N}{n}} = 1$$

para todo $m \leq M \leq N$.

Ahora, se observa que

$$E(n-m) = \sum_{n=m}^{N-M+m} \frac{(n-m)m \begin{pmatrix} M \\ m \end{pmatrix} \begin{pmatrix} N-M \\ n-m \end{pmatrix}}{n \begin{pmatrix} N \\ n \end{pmatrix}}$$

Y en esta suma se sustituyen las identidades:

$$\begin{pmatrix} M \\ m \end{pmatrix} = \frac{m+1}{M+1} \begin{pmatrix} M+1 \\ m+1 \end{pmatrix} = \frac{m^*}{M+1} \begin{pmatrix} M^* \\ m^* \end{pmatrix}$$

$$(n-m) \begin{pmatrix} N-M \\ n-m \end{pmatrix} = (N-M) \begin{pmatrix} N-M-1 \\ n-m-1 \end{pmatrix} = (N-M) \begin{pmatrix} N-M^* \\ n-m^* \end{pmatrix}$$

donde $M^* = M+1$ y $m^* = m+1$. Entonces,

$$E(n-m) = \frac{m(N-M)}{M+1} \underbrace{\sum_{n=m}^{N-M^*+m^*} \frac{m^* \begin{pmatrix} M^* \\ m^* \end{pmatrix} \begin{pmatrix} N-M^* \\ n-m^* \end{pmatrix}}{n \begin{pmatrix} N \\ n \end{pmatrix}}}_{=1} = \frac{m(N-M)}{M+1}$$

$$E(\hat{N}) = \frac{(M+1)}{m} E(n-m) + M = \frac{(M+1)}{m} \frac{m(N-M)}{M+1} + M = M$$

Eficiencia. El concepto de eficiencia está relacionado con el concepto de precisión del estimador. Primero se verá la definición de eficiencia relativa.

Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son ambos estimadores insesgados de θ, entonces si se satisface la desigualdad

$$V(\hat{\theta}_1) \leq V(\hat{\theta}_2)$$

diremos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$. Un estimador es más eficiente (más preciso), si su varianza es menor.

Un estimador θ es eficiente si $V(\hat{\theta} \leq V(\hat{\theta}^*)$ donde $\hat{\theta}^*$ es cualquier otro estimador.

Cuando se tiene un estimador insesgado del parámetro θ es útil saber si hay otro estimador insesgado más eficiente, en este sentido el teorema de Cramér-Rao proporciona una cota mínima para las varianzas de los estimadores insesgados. Aquí se presenta este teorema sin demostración.

TEOREMA 4.2 (Cramér Rao). *Si $X_1,\ X_2,\ X_3,\ \ldots,\ X_n$ es una muestra aleatoria de una función de densidad tal que $f(x;\ \theta) > 0$ para x en una región que no depende del parámetro entonces para todo estimador insesgado de θ se satisface la desigualdad*

$$V(\hat{\theta}) \geq \frac{1}{nE(\frac{\partial}{\partial\theta}\log(f(x;\ \theta)))^2}$$

Si se tiene un estimador que alcanza la cota de Cramér Rao, estaremos seguros que no hay otro estimador insesgado de θ que tenga menor varianza.

TEOREMA 4.3. *Un estimador insesgado $\hat{\theta}$ del parámetro θ es eficiente si alcanza la cota de Cramer-Rao.*

Consistencia. Si no es posible emplear estimadores de mínima varianza, el requisito mínimo deseable para un estimador, es que a medida que el tamaño de la muestra crezca, el valor del estimador tienda a estar cerca del valor del parámetro, propiedad que se denomina consistencia. Existen diversas definiciones de consistencia, más o menos restrictivas, pero la más utilizada es la siguiente.

Un estimador $\hat{\theta}$ del parámetro θ se dice que es consistente si y sólo si, cuando n crece, $\hat{\theta}$ tiende en probabilidad a θ. Esto significa que si para toda $\varepsilon > 0$ se tiene que si

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

entonces $\hat{\theta}$ es consistente.

La idea de la consistencia es que conforme el tamaño de la muestra aumenta, el valor del estimador se aproxima a la del parámetro con probabilidad igual a 1. Una condición necesaria para determinar que un estimador es consistente se presenta en el siguiente teorema

TEOREMA 4.4. *Si la varianza de un estimador insesgado es tal que*

$$\lim_{n\to\infty} V(\hat{\theta}) = 0$$

entonces $\hat{\theta}$ es un estimador consistente de θ.

Suficiencia. E1 concepto de estadística suficiente fue introducido por Fisher en 1922, y ha sido objeto de numerosas e importantes investigaciones. Como originalmente indicó Fisher, un estadístico (o función de la muestra de valores observados) es suficiente para el objetivo de la inferencia estadística si contiene, en un cierto sentido, toda la ‘‘información’’ sobre la distribución generadora (función de distribución de acuerdo con la cual ha sido generada la muestra de valores observados). ¿En qué sentido hemos de usar la palabra

"información" aquí? Supondremos que el modelo estadístico de las variables aleatorias observables tiene una cierta función de distribución conjunta que pertenece a una familia especificada $F(x;\ \theta)$ de funciones de distribución. Sin embargo, la verdadera función de distribución generadora es desconocida pues se desconoce el valor de θ.

Sea $X_1,\ X_2,\ \ldots,\ X_n$ una muestra aleatoria de la función de densidad $f(x;\ \theta)$; se llama estadística a cualquier función de los datos muestrales, esto es, cualquier función de la forma $T = h(X_1,\ X_2,\ \ldots,\ X_n)$ es una estadística.

La única información que se tiene para tomar una decisión sobre θ es el resultado del experimento aleatorio, es decir, la muestra aleatoria $X_1,\ X_2,\ \ldots,\ X_n$. Sin embargo, los datos muestrales son un complicado conjunto de números y el investigador se ve en la necesidad de introducir una simplificación deseable. Para esta simplificación, siempre que sea posible, se elegirá, un estadístico que pierda la menor información relativa al parámetro contenida en la muestra. Este es el deseo que incita a la definición de estadístico suficiente. Supóngase, pues, que T es una estadística (es decir, una función medible de la variable aleatoria X) y sea Z otra estadística; si consideramos la probabilidad condicionada de Z, dado T, en general esta probabilidad dependerá de θ; si ocurre que la función condicionada no depende del parámetro, querrá, decir que la estadística Z en presencia de T no proporciona ninguna información adicional acerca de θ. Si esto ocurre para cualquier otra estadística se concluye que toda la información que existía en la muestra $X_1,\ X_2,\ \ldots,\ X_n$ nos la ha proporcionado la estadística T, y en este caso se llama estadística suficiente.

Una estadística T se dice que es suficiente para el parámetro θ si la función de densidad condicional de U, cualquier otra estadística dada T no depende del parámetro. Esto es, si para toda U se tiene que $f_{U|T}(u|t)$ no depende de θ entonces T es estadística suficiente para θ.

En otras palabras, un estimador es suficiente para θ cuando ya tiene toda la información sobre el parámetro que está contenida en la muestra.

Es difícil probar con todas las posibles estadísticas si $f_{U|T}(u|t)$ no depende de θ, y debido a que todas las estadísticas dependen de la muestra, se puede tomar en lugar de todas las estadísticas la propia muestra $X_1,\ X_2,\ \ldots,\ X_n$. Entonces, un método para verificar si una estadística T es suficiente, es determinar la distribución condicionada de $X_1,\ X_2,\ \ldots,\ X_n$, dado T. Sin embargo, este método es también a menudo laborioso y de gran dificultad. Fisher y después Neyman proporcionaron un criterio simple, con el que podemos generalmente determinar si una familia $F(x;\ \theta)$ de funciones de distribución admite un estadístico suficiente no trivial y cuál es la forma de este estadístico.

Este criterio es proporcionado por el célebre teorema de factorización de Neyman-Fisher.

TEOREMA 4.5 (factorización de Neyman-Fisher). . *Sea T una estadística de la muestra aleatoria de la función de densidad $f(x;\ \theta)$, T es estadística suficiente del parámetro si y sólo si, la función de densidad conjunta se puede escribir como el producto de dos funciones, una que depende de la estadística suficiente y el parámetro y otra que depende únicamente de la muestra, esto es:*

$$f(x_1,\ x_2,\ \ldots,\ x_n;\ \theta) = f(x_1;\ \theta)f(x_2;\ \theta)\ldots f(x_n;\ \theta) = g(T;\ \theta)h(x_1,\ x_2,\ \ldots,\ x_n)$$

Robustez. Al estudiar un proceso aleatorio es posible que la función de distribución generadora de la muestra sea $F(x;\ \theta)$ (por ejemplo una exponencial) y se está suponiendo que la distribución correcta es $G(x;\ \theta)$, (por ejemplo una Weibull) y la estimación de θ no es afectada fuertemente por considerar que la distribución es $G(x;\ \theta)$ en lugar de $F(x;\ \theta)$, se dirá que el estimador es robusto.

Se dice que $\hat{\theta}$ es un estimador robusto del parámetro θ si los supuestos de partida en los que se basa la estimación, atribuida a la selección de la función de distribución que, en realidad, no es la correcta, no alteran de manera significativa los resultados que éste proporciona.

Método de Máxima Verosimilitud para generar Estimadores

El método de máxima verosimilitud es el método de estimación más ampliamente utilizado y se basa en escoger como estimador del parámetro al valor que maximiza la función de probabilidad dados los valores de la muestra.

Dada una muestra aleatoria $X_l,\ X_2,\ \ldots,\ X_n$, con función de densidad $f(x;\ \theta)$, se conoce como función de verosimilitud a la función de densidad conjunta de la muestra, en esta función el parámetro es la variable y la muestra se considera fija, esto es

$$L(\theta; x_1,\ x_2,\ x_3,\ \ldots,\ x_n) = f(x_1;\ \theta)f(x_2;\ \theta)f(x_3;\ \theta)\ldots f(x_n;\ \theta)\ \text{ con } \theta \in \Theta$$

Ya que las observaciones de una muestra aleatoria son independientes, la función de verosimilitud es el producto de las funciones de densidad marginales evaluada en los datos muestrales. Debido a que en la función de verosimilitud la variable es el parámetro y los valores $x_1,\ x_2,\ x_3,\ \ldots,\ x_n$ se consideran conocidos y en consecuencia, fijos, por simplicidad se utiliza la notación,

$$L(\theta) = L(\theta; x_1,\ x_2,\ x_3,\ \ldots,\ x_n) \quad \text{con } \theta \in \Theta$$

Dada una muestra aleatoria $X_1,\ X_2,\ \ldots,\ X_n$ con función de densidad $f(x;\ \theta)$, se conoce como estimador de máxima verosimilitud del parámetro θ al número donde la función de verosimilitud alcanza el valor máximo.

Entonces el objetivo del método de máxima verosimilitud es encontrar el valor dentro del posible rango de valores del parámetro que optimiza la función de verosimilitud, dados los datos de la muestra.

EJEMPLO 4.6. Encuentre el estimador de máxima verosimilitud del número de tanques que tiene el enemigo.

Para resolver este ejercicio primero se encuentra la función de verosimilitud de N dada la muestra $X_1,\ X_2,\ \ldots,\ X_n$.

Como los tanque en la muestra se obtienen sin reemplazo, una vez que se observa un número de un tanque, este ya no se vuelve a considerar para entrar en la muestra, además se supone que la selección es aleatoria y que cada tanque tiene la misma probabilidad de ser observado; en estas condiciones se tiene que

- La probabilidad que X_1 esté en la muestra es $1/N$,
- La probabilidad que X_2 esté en la muestra, dado que ya se seleccionó a X_1 es igual a $1/(N-1)$,
- La probabilidad que X_3 esté en la muestra, dado que ya se seleccionó a X_1 y a X_2 es igual a $1/(N-2)$, así se sigue hasta llegar a
- La probabilidad que X_n esté en la muestra, dado que ya se seleccionó a $X_1,\ X_2,\ \ldots,\ X_{n-1}$ es igual a $1/(N-n+1)$.

Entonces, la función de verosimilitud es igual a

$$L(N) = L(N;\ X_1,\ X_2,\ \ldots,\ X_n) = \frac{1}{N} \times \frac{1}{N-1} \times \frac{1}{N-2} \times \ldots \times \frac{1}{N-n+1}$$

Como se puede ver, la función de verosimilitud $L(N)$ no es continua, por lo que no se puede utilizar las técnicas del cálculo diferencial e integral para encontrar el valor de N que máximiza a $L(N)$. Pero haciendo un análisis de $L(N)$ se puede ver que es una función decreciente, y que $L(N) \to 0$ cuando $N \to \infty$. Por lo tanto $L(N)$ alcanza su valor más alto cuando N toma el valor más pequeño posible dada la muestra.

Por otro lado se sabe que $N \geq X_1$, $N \geq X_2$, ..., $N \geq X_n$, entonces $N \geq \max\{X_1,\ X_2,\ \ldots, X_n\}$, por lo que el valor máximo de $L(N)$ dada la muestra es cuando $N = \max\{X_1,\ X_2,\ \ldots, X_n\}$ y este es el estimador de máxima verosimilitud de N.

Ejercicios

1.- Probar que en el método de captura y recaptura el estimador $\hat{N}_4 = \frac{(n+1)(M+1)}{m+1} - 1$ es insesgado.

2.- Para $N = 100$, $M = 20$ y $n = 10$ encuentre los valores de $\hat{N}_4$ y sus probabilidades y gráfique estas probabilidades.

3.- Diga ?‘cuál de los tres estimadores de N, el número de tanques es el más eficiente.

4.- Diga ?‘cuál de los tres estimadores de N, el número de tanques es suficiente

5.- Diga ?‘cuál de los tres estimadores de N, el número de tanques es consistente.

Capítulo 5

Prueba de Hipótesis

Las pruebas de hipótesis es una parte de la inferencia estadística que consiste, básicamente, en decidir cual de dos posibles conjeturas sobre la población es verdadera. La decisión se hace con base a la información contenida en una muestra aleatoria.

Las conjeturas relativas a la población objeto de las pruebas de hipótesis pueden estar relacionadas ó bien con la forma de la distribución de una variable aleatoria (por ejemplo si la distribución generadora de los datos es una Weibull), ó con los valores de uno o varios parámetros de la misma (por ejemplo si el número de peces es $N = 200$).

De esta manera, las pruebas de hipótesis presentan dos enfoques:

- Pruebas de hipótesis sobre parámetros, consiste en determinar si el parámetro de una distribución toma o no un determinado valor, y
- Pruebas de Bondad de Ajuste, para definir si un conjunto de datos se ha generado de una determinada distribución.

En ambos casos, se debe tomar una decisión con base en los datos de una muestra aleatoria. Con el siguiente problema se pretende explicar los conceptos de la prueba de Hipótesis sobre un parámetro.

Problema 3.
Una pelea de Campeonato de Boxeo

En una pelea de box de campeonato hay un boxeador que tiene la corona de campeón mundial de boxeo, y hay otro boxeador que está retando al campeón la corona. El retador quiere demostrar que ahora el mejor boxeador del mundo es él.

□

En este contexto, al realizarse la pelea de box se tienen dos conjeturas:

- El campeón es el mejor boxeador.
- El retador es el mejor boxeador.

Estas dos conjeturas no tienen el mismo peso, pues se da por hecho que el mejor boxeador es el campeón mundial, por lo tanto, el retador debe probar contundentemente que el es mejor que el actual campeon. Esto significa que ante un empate, la decisión de los jueces es que el campeón sigue siendo el campeón. Esto significa que la conjetura de que el campeón es el mejor boxeador tiene ventaja sobre la conjetura el retador es el mejor boxeador, pues aún cuando no se realizara la pelea, la decisión sería reconocer que el mejor es el campeón.

La conjetura que tiene ventaja, porque es verdadera, según la historia del proceso o lo que establece el modelo utilizado, se le denota como hipótesis nula. La conjetura que se contrapone a la hipótesis nula y es la que se pone a prueba es la hipótesis alternativa. En este ejemplo, se tiene que

- La hipótesis nula es: ''El campeón es el mejor boxeador''.
- La hipótesis alternativa es: ''El retador es el mejor boxeador''.

La pelea se efectúa para tomar la decisión de cual de las dos conjeturas es la verdadera se pacta la pelea. Cada round de la pelea forma parte de la muestra, en base a lo observado en la muestra se tomara la decisión.

Los criterios para decidir, entre otros, son:

- La diferencia de golpes a favor del retador.
- El número de caidas.
- la fuerza de los golpes.

Los jueces de la pelea, tienen que tomar una decisión con base en lo que observen en cada uno de los rounds. Sin embargo, es claro que al desarrollarse la pelea interviene el azar, entonces la decisión no necesariamente es acorde con la realidad, pues eventos fortuitos están presentes durante la pelea.

La siguiente tabla nos muestra las posibles relaciones entre la realidad que desconocemos y la decisión que toman los jueces de la pelea.

	La realidad es	
La decisión declara que	**El campeón es mejor**	**El retador es mejor**
El campeón es mejor	Acierto	Error tipo II
El retador es mejor	Error tipo I	Acierto

Como se puede ver, los jueces pueden decidir que el campeón es mejor o pueden decidir que el retador es mejor y por lo tanto, que hay nuevo campeón, pero la realidad no se conoce. Si la decisión es que el campeón es mejor, y en realidad el campeón es mejor, entonces se habrá tomado una decisión adecuada, lo mismo ocurre si se decide que el retador es mejor y realmente el retador es mejor. Pero, si el campeón

es mejor y se decide que el retador es mejor, se habra cometido un error (error tipo I), lo mismo ocurre si se decide que el campeón es mejor cuando en realidad el retador es mejor, en este caso se comete el error tipo II.

Lo ideal sería tener un criterio de decisión con el cual la probabilidad de cometer cualquiera de los errores, tipo I o tipo II, fuera pequeña, sin embargo esto no es posible pues si relajamos las condiciones para aceptar la hipótesis nula, se endurecen las restricciones para rechazarla, y viceversa.

Elementos de la prueba de Hipótesis. Una hipótesis estadística es una afirmación o conjetura acerca de la función de distribución $F(x,\ \theta)$ generadora de la muestra aleatoria. Si la hipótesis estadística identifica por completo la distribución, recibe el nombre de "hipótesis simple" (por ejemplo que el número de tanques es $N = 100$) y si no la especifica recibe el nombre de "hipótesis compuesta" (por ejemplo que $N > 100$).

De esta manera, se tiene que

- Una hipótesis simple es de la forma: $\sigma^2 = 100$, $N = 16$, $p = 0.5$ ó "la función generadora de la muestra es una normal estándar"
- Una hipótesis compuestas es de la forma: $N > 100$, $\sigma^2 > 16$, $p < 0.5$, ó "la función generadora de la muestra es simétrica respecto al cero".

La prueba de hipótesis esencialmente consiste en decidir entre dos conjeturas opuestas cuál es la verdadera. Una de las conjeturas se supone verdadera ya sea porque la historia o la experiencia así lo ha establecido o porque así lo indica el modelo que genera los datos. La otra conjetura es la que los datos parecen respaldar. La primera conjetura se denota como hipótesis nula y la segunda conjetura se denota como hipótesis alternativa. Generalmente, la hipótesis alternativa es la hipótesis que defiende el investigador

Se llama hipótesis alternativa a la conjetura que se pone a prueba. Es la hipótesis de investigación. La hipótesis alternativa se denota como H_a ó H_1

Se llama hipótesis nula a la conjetura que se considera cierta, ya sea porque el modelo así lo indica, porque la historia lo ha probado o porque es lo aceptado. La hipótesis nula se identifica con el símbolo H_o

Se llama estadística de prueba a la función de los datos muestrales que se utiliza para tomar la decisión.

Se llama región crítica o región de rechazo para H_o al conjunto de valores de la estadística de prueba que hace que se rechace la hipótesis nula.

Dado que la decisión se hace con base en una muestra aleatoria es posible que la decisión que se tome sea equivocada, tanto si se rechaza la hipótesis nula, como si no se rechaza. La siguiente tabla muestra la relación entre la veracidad de las hipótesis de prueba y la decisión del investigador.

		La realidad	
		H_o **es verdadera**	H_o **es falsa**
La decisión es	**No rechazar** H_o	Acierto	Error tipo II
	Rechazar H_o	Error tipo I	Acierto

Como se puede ver en la tabla, hay dos tipos error.

El error tipo I se comete al rechazar la hipótesis nula cuando es verdadera. El error tipo II se comete al no rechazar la hipótesis nula cuando es falsa.

El nivel de significancia de la prueba de hipótesis es la probabilidad de cometer el error tipo I, y se denota con la letra α.

$$\text{P(error tipo I)} = \alpha = \text{nivel de significancia de la prueba}$$

La potencia de la prueba es la probabilidad de acertar cuando se rechaza la hipótesis nula, esto es, es la probabilidad de rechazar la hipótesis nula cuando es falsa. En particular, si el parámetro toma los valores de la hipótesis alternativa se tiene que:

$$1 - P(\text{error tipo II}) = \text{potencia de la prueba.}$$

Es deseable que la región de rechazo sea tal que, la probabilidad de cometer los errores tipo I y tipo II sean pequeñas; sin embargo, no es posible disminuir la probabilidad de ambos errores simultáneamente, pues conforme disminuye uno de ellos, aumenta el otro.

Problema 4:
Aplicación de un examen

Para ingresar a cualquier universidad generalmente se aplica un examen de opción múltiple, en la mayoría de los casos, este examen consiste de 120 preguntas con 5 diferentes opciones, señaladas como a, b, c, d, e. y una solamente de estas respuestas es la correcta. Para cada aspirante a ingresar a la universidad al presentar el examen, se formula una prueba de hipótesis, donde las conjeturas son:

(1) El aspirante tiene los conocimientos suficientes para ingresar a la universidad.
(2) El aspirante no tiene los conocimientos suficientes para ingresar a la universidad.

El aspirante debe probar que tiene los conocimientos suficientes, esta es la hipótesis alternativa. La hipótesis nula es que el aspirante no tiene los conocimientos suficientes para ingresar a la Universidad

- Hipótesis nula: *Ho*: El aspirante tiene los conocimientos suficientes para ingresar a la universidad.
- Hipótesis alternativa: *Ha*: El aspirante no tiene los conocimientos suficientes para ingresar a la universidad.

La calificación del examen es la estadística de prueba y se utiliza para decidir si el aspirante es aceptado o no.

Se rechaza la hipótesis nula cuando la calificación del examen es mayor o igual a 60.

El error tipo I, (rechazar Ho cuando es verdadera) es que el estudiante pase el examen sin que tenga los conocimientos suficientes para ingresar a la Universidad.

El error tipo II (aceptar Ho cuando es falsa) es que el estudiante repruebe el examen cuando si tiene los conocimientos suficientes para ingresar a la Universidad.

La hoja de respuesta del examen es de la forma

1. (a) (b) (c) (d) (e)	50. (a) (b) (c) (d) (e)	90. (a) (b) (c) (d) (e)
2. (a) (b) (c) (d) (e)	51. (a) (b) (c) (d) (e)	91. (a) (b) (c) (d) (e)
3. (a) (b) (c) (d) (e)	52. (a) (b) (c) (d) (e)	92. (a) (b) (c) (d) (e)
4. (a) (b) (c) (d) (e)	53. (a) (b) (c) (d) (e)	93. (a) (b) (c) (d) (e)
⋮	⋮	⋮

La probabilidad de elegir una respuesta correcta es p, conforme más preparado está el estudiante, mayor será la probabilidad de acertar en cada pregunta. La calificación del examen es una variable aleatoria cuya función de probabilidad es binomial con parámetros $n = 120$ y p. Si el aspirante no tiene ningún conocimiento y solo está contestando al azar, se tendrá que $p = 1/5 = 0.20$. Si el estudiante conoce el tema se tendrá que $p > 0.20$

Entonces las hipótesis de prueba se pueden escribir así,

$$Ho: \ p = 0.20 \quad \text{contra} \quad Ha: \ p > 0.20$$

$$\text{P(error tipo I) = P(que el estudiante sea admitido cuando no está preparado)}$$

$$= P(Calificación \geq 60 \mid p = 1/5) \simeq 0$$

Esta probabilidad se calcula considerando la función de probabilidad binomial.

Como se ve, es casí imposible que un estudiante que no sabe nada, pueda obtener más de 60 aciertos.

Si el aspirante tiene algún conocimiento, la probabilidad de acertar p es mayor a 0.20, pero cual es la potencia de la prueba para los diferentes valores de $p > 0.20$

La potencia de la prueba es

$$\begin{aligned} Potencia &= 1 - P(errorII) \\ &= 1 - P(Calificación < 60 | p > 0.20) \end{aligned}$$

En la siguiente figura se muestra la potencia de la prueba, en función de la probabilidad p.

□

Como se puede ver, conforme mayor sea la probabilidad de acertar, aumenta la potencia de la prueba.

Si se quiere disminuir la probabilidad de que alguien que no sabe pase el examen, se puede auemtar el puntaje de acreditación, esto significa que para pasar el examen se puede pedir que la calificación sea mayor o igual a 80. Es más difícil que alguien que no está preparado sea admitido, sin embargo aumenta la probabilidad que sea rechaado alguien que si está preparado. Esto significa que al disminuir la probabilidad de uno de los errores, aumenta la probabilidad del otro error.

En resumen: Una prueba o contraste de una hipótesis estadística es una regla o procedimiento que conduce a la decisión de aceptar o rechazar cierta hipótesis, identificada como hipótesis nula, con base en los resultados de una muestra. Los procedimientos de prueba de hipótesis dependen del empleo de la información contenida en una muestra aleatoria de la población de interés. Si esta información es consistente con la hipótesis nula se concluye que esta es verdadera; sin embargo, si esta información es inconsistente con la hipótesis nula se concluye que esta es falsa.

Los pasos a seguir en una prueba de hipótesis es:

(1) Formular las hipótesis.
(2) Tomar una muestra aleatoria de la variable de interés $X_1, X_2, \ldots, X_n$.
(3) Generar o calcular un "Estadístico de prueba", que servirá para definir la acción de aceptar o rechazar la hipótesis nula.
(4) Definir el criterio de aceptación o de rechazo. Es decir, el procedimiento de prueba parte los posibles valores del estadístico de prueba en dos subconjuntos o regiones: Una "región de aceptación de H_0" y una "región de rechazo de H_0".

(5) Tomar la decisión de aceptar o rechazar H_0 dependiendo de si el estadístico de prueba queda en la región de aceptación o en la región de rechazo.

Es importante comprender que la aceptación de una hipótesis nula simplemente implica que los datos obtenidos no dan suficiente evidencia para rechazarla. Por otro lado, el rechazo de una hipótesis implica que la evidencia muestral pone en duda la hipótesis planteada.

El lema de Neuman Pearson proporciona una forma de obtener "la mejor" región de rechazo, esto es, la región de rechazo que tiene menor probabilidad de error II, fijando la probabilidad del error I.

TEOREMA 5.1 (Lema de Neuman Pearson). *Dadas las hipótesis simples*

$$H_0 : \ \theta = \theta_0 \quad \text{contra} \quad H_1 : \ \theta = \theta_1$$

la región dada por

$$C = \{X_1, \ X_2, \ \ldots, \ X_n \mid \frac{L(\theta_0}{L(\theta_1)} < \lambda\}$$

es la mejor región crítica, en el sentido que

- $P(errorI) = P(C \mid \theta = \theta_0) = \alpha$ *y*
- $P(errorII) = P(C \mid \theta = \theta_1) = \beta$ *con* β *mínima.*

Ejercicio

1.- Sea $X_1, \ X_2, \ X_3, \ \ldots, \ X_n$ una muestra aleatoria Bernoulli con parámetro p, $(f(x) = p^x(1-p)^{1-x}$, para $x = 0, \ 1)$. Encuentre la mejor región crítica para las hipótesis

$$H_0 : \ p = 0.4 \quad \text{contra} \quad H_a : \ p = 0.8$$

2.- Sea $X_1, \ X_2, \ X_3, \ \ldots, \ X_n$ una muestra aleatoria Poisson con parámetro p, $(f(x) = \frac{\lambda^x e^{-\lambda}}{x!})$. Encuentre la mejor región crítica para las hipótesis

$$H_0 : \ \lambda = 5 \quad \text{contra} \quad H_a : \ \lambda = 2$$

Capítulo 6

Análisis de Regresión lineal

El análisis de regresión lineal es una técnica estadística que estudia la relación funcional entre una variable de interés y una o más variables explicativas. Esto es: si Y es la variable de interés, y X_1, X_2, ..., X_m son las variables explicativas, se supone que existe una relación entre Y y X_1, X_2, ..., X_M de la forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

donde $\varepsilon \sim N(0, \sigma)$ y β_0, β_1, ..., β_m y σ son parámetros desconocidos.

En este modelo se pueden estimar los parámetros o se puede hacer prueba de hipótesis sobre los mismos. En este capítulo se presenta un problema que se puede modelar con una regresión lineal.

Problema 5:
La producción de fruta en una huerta

Se quiere instalar una huerta de árboles frutales y se quiere determinar cuántos árboles deben ser sembrados para que se tenga una producción promedio máxima cuando llegue a ser productiva. Se sabe que si se siembran más árboles, la producción total tendrá más contribuyentes, pero por la competencia del suelo, cada árbol disminuye un poco su producción individual. Esto es, un único árbol sembrado en la huerta produce a kilogramos de fruta en promedio al año, dos árboles sembrados en la huerta producen cada uno $a - b$ kilogramos de fruta en promedio y la producción total de la huerta es $2(a - b)$. Por cada nuevo árbol sembrado, aumenta la producción total de la huerta porque hay nuevos árboles que contribuyen en la suma total, pero la producción individual de cada árbol disminuye en b kilogramos en promedio al año porque aumenta la competencia por los nutrientes del suelo. Lo que se busca en este ejemplo es conocer el número de árboles que dan la mayor producción anual.

En la siguiente tabla se describe la dinámica de la producción promedio anual de la huerta, en función de los árboles sembrados.

Número de árboles en la huerta	Producción esperada de cada árbol	Producción esperada de la huerta
1	a	a
2	$a-b$	$2(a-b)$
3	$a-2b$	$3(a-2b)$
4	$a-3b$	$4(a-3b)$
$\vdots$	$\vdots$	$\vdots$
x	$a-(x-1)b$	$x(a-(x-1)b)$
$\vdots$	$\vdots$	$\vdots$

Esto significa que la producción promedio de la huerta al año sigue la ecuación

$$E(Y \mid x) = x(a-(x-1)b) = x((a+b)-bx) = (a+b)x - bx^2,$$

donde a y b son dos parámetros desconocidos.

Esta ecuación no está totalmente determinada pues se desconoce el valor de a y de b, por lo que estos dos valores se deben estimar. Para hacer la estimación se utilizan los datos de la producción del año pasado de 24 huertas de igual condición y superficie.

Los datos observados se encuentran en la siguiente tabla, X es el número de árboles en la huerta y Y es el número de kilogramos de fruta en la producción anual.

X	60	65	70	75	80	85	90	95	100	105	110
Y	1605	1803	1890	1842	1963	1990	1931	2034	2056	2006	1960
X	120	125	130	135	140	145	150	155	160	165	170
Y	1885	1890	1780	1770	1630	1610	1440	1397	1280	1100	1017

Las observaciones son datos aleatorios y se supone que siguen el modelo lineal

$$Y = \beta_1 x + \beta_2 x^2 + \varepsilon$$

con ε una variable aleatoria tal que $\varepsilon \sim N(0, \sigma^2)$, y $\beta_1 = a + b$, $\beta_2 = -b$ y σ^2 son constantes desconocidas. Con los datos se pueden estimar los parámetros β_1, β_2 y σ^2.

El método de estimación aplicado es el de mínimos cuadrados. La idea de este método de estimación es encontrar la ecuación que se aproxime más a los datos, esto significa encontrar los valores de β_1 y β_2 que minimice la suma de los errores aleatorios al cuadrado, esto es, se resuelve el problema

Minimizar la suma

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} \left(Y_i - \beta_1 x - \beta_2 x^2\right)^2$$

Para resolver este problema se deriva la ecuación respecto a β_1 y respecto a β_2, luego se iguala a 0 cada derivada y se resuelve el sistema de ecuaciones lineales.

$$\begin{aligned} \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \varepsilon_i &= 2 \sum_{i=1}^{n} \left(Y_i - \beta_1 x_i - \beta_2 x_i^2\right) x_i \\ &= 2 \left(\sum_{i=1}^{n} Y_i x_i - \beta_1 \sum_{i=1}^{n} x_i^2 - \beta_2 x_i^3\right) \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \beta_2} \sum_{i=1}^{n} \varepsilon_i &= 2 \sum_{i=1}^{n} \left(Y_i - \beta_1 x_i - \beta_2 x_i^2\right) x_i^2 \\ &= 2 \left(\sum_{i=1}^{n} Y_i x_i^2 - \beta_1 \sum_{i=1}^{n} x_i^3 - \beta_2 x_i^4\right) \end{aligned}$$

El sistema de ecuaciones lineales es

$$\begin{aligned} \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} Y_i x_i \\ \beta_1 \sum_{i=1}^{n} x_i^3 + \beta_2 \sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} Y_i x_i^2 \end{aligned}$$

De esta manera se tiene que los estimadores son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i^4 \sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n} x_i^3 \sum_{i=1}^{n} Y_i x_i^2}{\sum_{i=1}^{n} x_i^4 \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^3\right)^2}$$

y

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} Y_i x_i^2 - \sum_{i=1}^{n} x_i^3 \sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^4 \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^3\right)^2}$$

En la siguiente tabla se muestran los cálculos requeridos para obtener los estimadores de β_1 y β_2.

X	Y	x^2	x^3	x^4	Yx	Yx^2	
60	1605	3600	216000	12960000	96300	5778000	
65	1803	4225	274625	17850625	117195	7617675	
70	1890	4900	343000	24010000	132300	9261000	
75	1842	5625	421875	31640625	138150	10361250	
80	1963	6400	512000	40960000	157040	12563200	
85	1990	7225	614125	52200625	169150	14377750	
90	1931	8100	729000	65610000	173790	15641100	
95	2034	9025	857375	81450625	193230	18356850	
100	2056	10000	1000000	100000000	205600	20560000	
105	2006	11025	1157625	121550625	210630	22116150	
110	1960	12100	1331000	146410000	215600	23716000	
115	1950	13225	1520875	174900625	224250	25788750	
120	1885	14400	1728000	207360000	226200	27144000	
125	1890	15625	1953125	244140625	236250	29531250	
130	1780	16900	2197000	285610000	231400	30082000	
135	1770	18225	2460375	332150625	238950	32258250	
140	1630	19600	2744000	384160000	228200	31948000	
145	1610	21025	3048625	442050625	233450	33850250	
150	1440	22500	3375000	506250000	216000	32400000	
155	1397	24025	3723875	577200625	216535	33562925	
160	1280	25600	4096000	655360000	204800	32768000	
165	1100	27225	4492125	741200625	181500	29947500	
170	1017	28900	4913000	835210000	172890	29391300	
175	812	30625	5359375	937890625	142100	24867500	
2820	40641	360100	49068000	7018127500	4561510	553888700	**Sum**

Sustituyendo estos valores en las fórmulas anteriores se sigue que

- $\hat{\beta}_1 = 40.441$ y
- $\hat{\beta}_2 = -0.2038$

De esta manera la ecuación estimada esta dada por:

$$\hat{Y} = 40.441x - 0.2038x^2$$

Ya estimamos la ecuación de producción, ahora se puede estimar el número de árboles donde se obtiene una producción óptima.

La derivada de $\hat{Y}$ con respecto a x es $\frac{d\hat{y}}{dx} = \hat{\beta}_1 + 2\hat{\beta}_2 x$ de aquí se sigue que la raíz de esta derivada es

$$\hat{\beta}_1 + 2\hat{\beta}_2 x = 0 \;\Rightarrow\; x = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$$

$$\hat{x}_{op} = 99.20$$

Lo que significa que en la huerta se debe sembrar de 98 o 99 árboles, para tener una producción máxima anual.

En la siguiente figura se muestran los datos y la gráfica de la función estimada. Como se puede ver, la ecuación estimada se ajusta bien a los datos.

□

Para determinar que tan bueno es el modelo propuesto para describir la tendencia de Y se utiliza el coeficiente de determinación o se puede efectuar una prueba de hipótesis sobre los parámetros del modelo.

Coeficiente de Determinación

El coeficiente de determinación es un índice que indica el porcentaje de variación explicado por el modelo propuesto. Para calcular al coeficiente de determinación, primero se obtienen tres sumas de cuadrados.

- Suma de cuadrados total.

  Esta suma mide la variación total de la variable Y y se define como:

$$SCt = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

  Observe que los sumandos en esta suma son los cuadrados de las desviaciones de cada valor de la variable.
- Suma de cuadrados de la regresión.

  Esta suma mide la variación de los estimadores de Y, y se define como:

$$SCR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

  Observe que los sumandos en esta suma son los cuadrados de las desviaciones de cada valor estimado de la variable.
- Suma de cuadrados del error.

  Esta suma mide la variación aleatoria y se calcula con la fórmula

$$SCE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  Observe que los sumandos son la diferencia entre el dato observado y el dato estimado. Esto corresponde a la magnitud de la separación del valor observado y la ecuación estimada.

Un resultado importante que aquí se presenta sin demostración es el siguiente:

TEOREMA 6.1. *Sean SCE, SCR y SCt las sumas de cuadrados definidas anteriormente, entonces se satisface la relación*

$$SCt = SCR + SCE \tag{6.1}$$

Este resultado es importante, pues permite determinar cuando la variación debida a la regresión o la variación aleatoria es pequeña, pues se compara con la variación total. Observe que mientras la SCE es más pequeña el ajuste de los datos a la ecuación propuesta es mejor.

Sean SCR y SCt, las sumas de cuadrados definidas anteriormente, se define el coeficiente de determinación como el cociente

$$R^2 = \frac{SCR}{SCt}$$

considerando la ecuación 6.1 se sigue que $0 \leq R^2 \leq 1$,

Cuando R^2 se encuentre cerca del 1, se tiene que el ajuste es muy bueno, porque la separación de los datos a la ecuación estimada es pequeña. Cuando R^2 se encuentra cerca de 0, implica que la variación aleatoria es grande o que la variable Y no depende de la variable explicativa.

X	Y	$\hat{Y}$	$(Y-\overline{Y})^2$	$(\hat{Y}-\overline{Y})^2$	
60	1605	1700.5641	7810.140625	51.68315881	
65	1803	1773.6576	12017.64063	6445.295863	
70	1890	1836.7061	38661.39063	20543.80423	
75	1842	1889.7096	22089.39063	38547.27516	
80	1963	1932.6681	72697.64063	57261.18771	
85	1990	1965.5816	87986.39063	74096.43308	
90	1931	1988.4501	56465.64063	87069.31464	
95	2034	2001.2736	116025.3906	94801.54788	
100	2056	2004.0521	131496.8906	96520.26046	
105	2006	1996.7856	97734.39063	92057.99219	
110	1960	1979.4741	71088.89063	81852.69502	
115	1950	1952.1176	65856.39063	66947.73305	
120	1885	1914.7161	36720.14063	48991.88255	
125	1890	1867.2696	38661.39063	30239.33191	
130	1780	1809.7781	7503.890625	13549.68169	
135	1770	1742.2416	5871.390625	2387.944596	
140	1630	1664.6601	4016.390625	824.545482	
145	1610	1577.0336	6951.390625	13535.32135	
150	1440	1479.3621	64198.89063	45801.52137	
155	1397	1371.6456	87838.14063	103509.8068	
160	1280	1253.8841	170878.8906	193152.2512	
165	1100	1126.0776	352093.8906	321826.34	
170	1017	988.2261	457483.1406	497234.9712	
175	812	840.3296	776821.8906	727686.4545	
	40641	**40656.2644**	**2788969.625**	**2714935.275**	**Suma**

Con los resultados de la tabla se calcula el coeficiente de determinación de estos datos, $R^2 = 0.9735$, esto significa que el 97.35% de la variación total de los datos es explicada por el modelo. Esto se puede interpretar diciendo que el ajuste es muy bueno, ya que el porcentaje de variación debida al azar es 2.65%

Prueba de la Regresión

La prueba de la regresión se formula así:

$$H_0: \ \beta_1 = 0 \ \text{ y } \ \beta_2 = 0 \quad \text{contra} \ \ \beta_1 \neq 0 \ \text{ ó } \ \beta_2 \neq 0$$

La hipótesis nula indica que Y no depende ni de X ni de X^2. La hipótesis alternativa indica que Y depende al menos de una de las dos variables, X ó X^2.

Para tomar la decisión se comparan las variaciones debidas al azar y al modelo de regresión. La estadística de prueba es $Fc = \frac{CMR}{CME}$ donde

$CMR = SCR/1$ y $CME = SCE/(n-2)$. Se rechaza la hipótesis nula cuando F_C es mayor que el valor de F al nivel de significancia α.

EJEMPLO 6.2. Encuentre la F_c con los datos de la huerta.

Para calcular la estadística de prueba primero se calcula

$$SCE = SCt - SCR = 2788969.625 - 2714935.275 = 74034.35$$

Y los cuadrados medios del error son: $CME = 74034.35/22 = 3365.197727$

La estadística de prueba es

$$F_c = \frac{CMR}{CME} = \frac{2714935.275}{3365.197727} = 806.7684264$$

El valor de tablas para $\alpha = 0.05$ es $F_\alpha = 4.3$

Dado que $F_c > F_\alpha$ se rechaza la hipótesis nula.

Referencias

Keeping E. S. ‘‘Introduction to statistical Inference’’, Dover, 1995.

Quevedo H. Pérez-Salvador B. R. ‘‘Estadística para Ingeniería y Ciencias’’, Ed. Patria, 2008